Skolkovo Institute of Science and Technology

# EVOLUTIONARY ANALYSIS OF INTRAHOST INTERACTION BETWEEN PATHOGENS AND ADAPTIVE IMMUNITY

Doctoral Thesis

by

EVGENIIA ALEKSEEVA

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor

Professor Georgii Bazykin

Moscow — 2023

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate (Evgeniia Alekseeva)

Supervisor (Prof. Georgii Bazykin)

# ABSTRACT

Intrahost interaction between pathogens and adaptive immunity produce unique scenarios when evolution may be studied and tracked in real time. In this context the power of evolutionary and phylogenetic analysis, mostly developed to study evolution on much larger scales, can be applied to objects unusual for classical evolutionary biology. Development of immune repertoire sequencing made possible to observe diversification of individual B cell lineages, occurred due to the process of affinity maturation. Intrahost evolution of pathogens is more frequently studied, however still attracts growing interest of the scientific community. Thus, such studies of intrahost interaction are in intersection of several biology fields, including immunology, virology, evolutionary and population genomics, and require understanding of observed processes from different sides. Moreover, frequently in such works rethinking and modification of classical evolutionary genomics approaches are required, which also make them methodologically interesting. Therefore, when combined, these factors contribute to the growth, development, practical importance, and immense interest in the relatively small field of intrahost evolutionary analysis.

This thesis joins together two works, which both include evolutionary analysis of short-scale intrahost interaction between the adaptive immunity and the pathogen. However these two studies consider this interaction from opposite sides of the host-pathogen arms race. In the first part (Chapter 3), we studied how B cell immunity may adapt to rapidly changing pathogens using affinity maturation of B cell clonal lineages. To do so we analyzed longitudinal B cell repertoires, sampled from peripheral blood of healthy volunteers three times within a year. Peripheral blood was sorted on cell subsets of memory B cells, plasmablasts and plasma cells, so observed B cell clonal lineages reflected both temporal and phenotypic dynamics of the lineage. We revealed two functionally different types of B cell clonal lineages: the first type, belonging to persisting memory and the second type of antibody-secreting lineages, involved in ongoing immune response. Difference in lineage functioning resulted in different modes of selection, shaping their evolution. Phylogenetic analysis provided evidence that these two functional states may transit between each other depending on the antigen challenge, showing memory reactivation followed by new cycles of affinity maturation.

The second part of the thesis (Chapter 4) investigates a case of long-term COVID-19 in an immunocompromised host. In contrast there we tracked intrahost evolution of the pathogen and studied the selection forces shaping its divergence from the initial virus. We revealed that almost a third of SARS-CoV-2 changes prevent or reduce binding of known immunogenic CD8 T cell epitopes to a patient's HLA allele. To the best of our knowledge this is the first evidence that the T cell escape can be a driver of intra-host evolution of SARS-CoV-2.

Therefore, findings of this work shows that evolutionary analysis of host-pathogen interaction is a powerful approach, which may shed light on various aspects of the functioning of the B cell adaptive immunity and dynamics of pathological conditions.

**Key words:** Rep-Seq, immune repertoires, affinity maturation, B cell clonal evolution, intrahost evolution, SARS-CoV-2, T cell immune escape

## PUBLICATIONS

# Papers

1. Mikelov AI*, **Alekseeva EI***, Komech EA, Staroverov DB, Turchaninova MA, Shugay M, Chudakov DM, Bazykin GA, Zvyagin IV. **Memory persistence and differentiation into antibody-secreting cells accompanied by positive selection in longitudinal BCR repertoires**. Elife. 2022 Sep 15;11:e79254. doi: 10.7554/eLife.79254. Epub ahead of print. PMID: 36107479.

2. Stanevich O*, **Alekseeva E***, Sergeeva M, Fadeev A, Komissarova K, Ivanova A, Simakova T, Vasilyev K, Shurygina A.-P, Stukova M, Safina K, Nabieva E, Garushyants S, Klink G, Bakin E, Zabutova J, Kholodnaia A, Lukina O, Skorokhod I, Ryabchikova V, Medvedeva N, Lioznov D, Danilenko D, Chudakov D, Komissarov A, Bazykin G. **SARS-CoV-2 escape from cytotoxic T cells during long-term COVID-19**. Nat Commun 14, 149 (2023). https://doi.org/10.1038/s41467-022-34033-x

3. Mukhina OA, Fomina DS, Parshin VV, Gushchin VA, Dolzhikova IV, Shchetinin AM, Chudakov DM, **Alekseeva EI**, Korostin D, Bazykin GA, Klink G, Logunov DY, Lysenko MA. **SARS-CoV-2 evolution in a patient with secondary B-cell immunodeficiency: A clinical case.** Human Vaccines & Immunotherapeutics. 2022 Aug 01; doi: 10.1080/21645515.2022.2101334. PMID: 35914217

4. Klink GV, Safina KR, Nabieva E, Shvyrev N, Garushyants S, **Alekseeva E**, Komissarov AB, Danilenko DM, Pochtovyi AA, Divisenko EV, Vasilchenko LA, Shidlovskaya EV, Kuznetsova NA; Coronavirus Russian Genetics Initiative (CoRGI) Consortium, Speranskaya AS, Samoilov AE, Neverov AD, Popova AV, Fedonin GG; CRIE Consortium, Akimkin VG, Lioznov D, Gushchin VA, Shchur V, Bazykin GA. **The rise and spread of the SARS-CoV-2 AY.122 lineage in Russia.** Virus Evol. 2022 Mar 5;8(1):veac017. doi: 10.1093/ve/veac017. PMID: 35371558

* equal contribution

# Conference presentations

1. Oral presentation: **Evolution of SARS-CoV-2 leads to T cell escape under the condition of persisting infection and immune depleted therapy**, Academy of Laboratory Medicine: advanced achievement, I Russian Congress with International Participation, July 2021, Moscow Russian (in person participation);

2. Poster presentation: **Signs of positive selection in new rounds of affinity maturation during B cell memory reactivation**, European Academy of Allergy and Clinical Immunology Hybrid Congress, July 2021, Krakow, Poland (online participation);

3. Poster presentation: **SARS-CoV-2 escape from cytotoxic T cells during long-term COVID-19**, International Moscow Conference on Computational Molecular Biology, July 2021, Moscow, Russia (in person participation);

# Awards

1. Grant RFBR, project number 19-31-27001 (Аспиранты 2020).

# ACKNOWLEDGEMENTS

This work would be impossible without the help and support of a lot of people. First of all I am deeply grateful to my supervisor Yegor Bazykin, who gave me the opportunity to spend my PhD years in his group. I am very thankful for Yegor for sharing his expertise and experience, giving freedom in my scientific search and in a way of thinking, friendly work atmosphere and human support in various situations.

Behind a cool chef there is a cool lab. I also want to thank all Yegor's group, which is a very friendly and warm team, where you always feel a pleasure to be. It is not just a group of nice people, but a team with a high-level of scientific integrity, where you always can get honest feedback on your work, find inspiration to try something new and ask new scientific questions. I started listing lab members, which I want to thank the most, but realized that I was just repeating the list of the laboratory from the website. So thank you all, guys, you are the best! I really hope that all of us will be safe and in touch and a lot of interesting projects and collaborations are waiting for us to be done together in spite of growing distances and circumstances, turned our lives around this year.

I am grateful to my co-authors for our work together, our discussions, shared ideas and experience. For the project of B cell clonal evolution I especially want to thank Artem Mikelov and Ivan Zvyagin, who designed the study, got the data and were open minded about various ideas, how to do different "evolutionary stuff" on their dataset. I am happy that these computational experiments finally resulted in a beautiful story about the work of B cell immunity. For the case study on long-term COVID-19, I would like to express my special gratitude to Anna-Polina Shurygina and Maria Sergeeva, who validated my computational conclusions through laboratory experiments. I am also very grateful for Mikhail Shugay, Alexey Neverov and Dmitrii Chudakov, who gave their valuable feedback and advice during my analysis in both projects.

All this work would be impossible without Aleksandr Tashkeev, who fascinated me to study bioinformatics together and encouraged me to enter the MS program in Skoltech. I am very grateful for him for the support, inspiration and a lot of help on this way. I also want to thank Yaroslav Ispolatov, my first supervisor at Skoltech, who guided me in my first computational steps and in general greatly expanded my worldview. I am very grateful for Skoltech in general for its community and the high level of science they are doing. And of course I am very grateful for my beloved family and friends for the huge amount of love and support they gave me in this way.

# TABLE OF CONTENTS

# LIST OF ABBREVIATIONS

ACE2 - angiotensin-converting enzyme 2;

AID - activation-induced cytidine deaminase;

AIRR - Adaptive Immune Receptor Repertoire;

AM - affinity maturation;

ART - antiretroviral therapy;

ASC - antibody-secreting cells;

BCR - B cell receptor;

Bmem - memory B cells;

BR - best rank;

CCP - convalescent plasma treatment;

CD - cluster determination molecule;

CDR - complementarity-determining region;

CLL - chronic lymphocytic leukemia;

DNA - deoxyribonucleic acid;

DNN - distances to the nearest neighbor;

DNA-PK - DNA-dependent protein kinase;

dsRNA - double stranded RNA;

FDC - follicular dendritic cell;

FL - follicular lymphoma;

FWR - framework region;

GC - germinal center;

gDNA - genomic DNA;

HIV - human immunodeficiency virus;

HLA - human leukocyte antigen;

HSCT - haematopoietic stem cell transplantation;

Ig - immunoglobulin;

LCMV - lymphocytic choriomeningitis virus;

MERS-CoV - Middle East respiratory syndrome coronavirus;

MHC - major histocompatibility complex;

MIG - molecular identifiers groups;

MRCA - most recent common ancestor;

mRNA - messenger RNA;

PBL - plasmablasts;

PHBR - patient harmonic best rank;

PL - plasma cells;

PRR - pattern recognition receptors;

RAG - recombination-activating gene;

RBD - Receptor Binding Domain;

Rep-Seq - repertoire sequencing;

RNA - ribonucleic acid;

RSS - recombination signal sequences;

RT - reverse transcription;

SARS-CoV-2 — severe acute respiratory syndrome coronavirus 2;

SFS - site frequency spectrum;

SHM - somatic hypermutations;

TdT - terminal deoxynucleotidyl transferase;

$T_{fh}$ - T follicular helper cell;

$T_h$ cell - T helper cell;

UMI - unique molecular identifiers;

VDJ segments - variable (V), diversity (D) and joining (J) segments;

VOC - variants of concern;

VOI - variants of interest.

# CHAPTER 1: INTRODUCTION

The arms race of host-pathogen coevolution has been going on since the origin of life (Tenthorey, Emerman, and Malik 2022). It is often described by the Red-Queen effect, named after the famous state of the Red Queen to Alice from Lewis Carroll's '*Through the Looking-Glass*': "It takes all the running you can do, to keep in the same place." (Carroll 1871). Both prokaryotes and eukaryotes have evolved defense mechanisms to secure cell integrity (Dunin-Horkawicz, Kopec, and Lupas 2014; Desjardins, Houde, and Gagnon 2005). With the development of multicellularity more than 600 million years ago specialized cellular lines (amebocytes, hemocytes, coelomocytes) have diverged, aimed to fight microbes by phagocytosis (Desjardins, Houde, and Gagnon 2005; Buchmann 2014). However, the trickiest part of defense mechanisms in multicellular organisms is not how to fight pathogens, but how to distinguish them from self cells and self intracellular structures. Thus the divergence of specialized phagocytes was accompanied by the development of the whole system of pattern recognition receptors (PRR), able to distinguish self from nonself by conservative features of pathogens such as elements of bacterial or fungi cell walls or nucleic acids, unusual for eukaryotic cells (dsRNA) (Buchmann 2014). Together with PRR the system of inner effector peptides and proteins were developed, aimed to fight pathogens by themselves (antimicrobial peptides, fibrinogen-related peptides, proteins of complement system) or orchestrate regulation of immune cells (chemokines) (Fujita 2002; Emery, Dimos, and Mydlarz 2021; Buchmann 2014). All mentioned innovations of immune defense belong to the branch of innate immunity, which possess effective weapons to fight against nonself structures and use the most conservative features of pathogens to distinguish between self and nonself molecules.

Concurrently pathogens didn't stand still either. They developed mechanisms to ease the invasion to the host and escape or manipulate host immune response (Weitz et al. 2019; Lord and Bonsall 2021). In addition in this war pathogens have one specific advantage. Usually generation time of pathogens is incomparably less than the generation time of the host. Therefore in general pathogens may modify their recognition patterns much faster than the host may develop new PRRs. Moreover pathogens have an opportunity to evolve right during contact with the host. Host defense mechanisms by itself positively select those variants of pathogens, which are harder to recognize. Such inequality in generation times and rate of evolution was compensated by the development of adaptive immunity 500-600 million years ago. In some literature this moment was called "Immunological Big Bang" (Sirisinha 2014; Flajnik and Kasahara 2010). Adaptive immunity is based on an absolutely innovative idea of pattern recognition. While innate immunity develops new PRRs to enlarge the number of patterns, it can recognize (Q. Zhang, Zmasek, and Godzik 2010), adaptive immunity aims to recognize any nonself molecule by generating a huge diversity of immune receptors with the deletion of receptors with self-recognition. The idea of adaptive immune receptors was independently developed several times in the evolution of vertebrates. Mechanisms of generation of large diversity of immune receptors can be roughly divided on RAG-independent, realized in jawless fishes (Boehm 2011; Boehm, Iwanami, and Hess 2012), and RAG-dependent that is characteristic of most jawed vertebrates (Cooper and Alder 2006). Anyway, large diversity of immune receptors may respond promptly to changes of pathogens by activation of new lymphocytes with corresponding receptors.

Hence, an arms race between host adaptive immunity and pathogens may be observed on a microevolutionary scale. Pathogens may evolve right inside the host, trying to avoid recognition by specific lymphocytes, constituting the adaptive immune response. Such alteration of pathogen escape and involvement of new immune receptors was observed for several viruses such as HIV, hepatitis and influenza. Moreover, the system of affinity maturation was developed in B lymphocytes (McCarthy et al. 2019; Muecksch et al. 2021). In this process B cell receptors can be additionally modified for high-affinity recognition of the antigen. In some cases, such in HIV infected individuals, maturation of B cell receptors occurs throughout the whole infection period following intrahost evolution of the pathogen (Nourmohammad et al. 2019; Bonsignori et al. 2017).

Thus host-pathogen coevolution may be observed and studied inside a single host on a short time scale. The goal of this thesis is to study mechanisms of intrahost evolution of adaptive immunity and pathogens using the power of approach of evolutionary genomics. To do this two different study designs and datasets were used:

1. The first part of the thesis (Chapter 3) is devoted to the study of evolutionary dynamics of affinity maturation in immune B cell repertoires. For this study peripheral blood samples were taken from four healthy individuals three times within a year. B cells were sorted to

three subsets: memory B cells, plasmablasts and plasma cells, and then full length sequencing of immunoglobulin heavy chains was performed. For evolutionary and phylogenetic analysis we used the most abundant B cell clonal lineages from joined repertoires of all B cell subsets and time points, which contain at least 20 unique B cell sequences. My contribution to this study is in the analysis of most abundant B cell lineages of these repertoires with the focus on association between dynamics and composition of B cell lineages and evolutionary regimes underlying their development. For this goal I adapted common population genetics approaches to the specific features of B cell clonal evolution and developed a pipeline, which can be used for evolutionary analysis of B cell immune repertoires.

2. In the second part of the thesis (Chapter 4) coevolution is studied from the side of the pathogen. There we analyze a case of long-term intrahost SARS-CoV-2 evolution in a patient with non-Hodgkin lymphoma under rituximab therapy without therapy by convalescent plasma. The patient has almost no B cell immune response and viral infection is mostly restrained by cytotoxic T cells. Using phylogenetic analysis of viral genome sequences we tracked how virus accumulated mutations, which prevent binding of viral antigens to the patient's HLA alleles. In this work I developed and performed an analysis of mutations effect on antigen presentation both at individual and population levels. As a result I predicted those mutations, which helped the virus to escape cytotoxic T cells, which then were experimentally validated by my coauthors. I also analyzed the effect of the patient's SARS-CoV-2 variant on CD8 T cell response in a general population.

In summary, my thesis joins together two different but complementary studies, showing intrahost evolution of B cell clonal lineages as a part of adaptive immunity and intrahost evolution of SARS-CoV-2 escaping cytotoxic T cells. It uses classical approaches of evolutionary and population genomics in analysis of evolutionary scenarios, happening before our eyes at small time scales.

# CHAPTER 2: LITERATURE REVIEW

## Principles of adaptive immunity

### Antigen receptors of T and B lymphocytes

The main players of adaptive immunity are T and B lymphocytes, which together possess a huge diversity of highly specific antigen receptors (Murphy and Weaver 2016; Kapila 2004). Antigen receptors of T and B lymphocytes are close in their structures and have a common evolutionary origin, however differ in their role in the immune response (Marchalonis, Jensen, and Schluter 2002). T cell receptors (TCRs) recognize short peptides of processed antigens, bound to the major histocompatibility complexes (MHC) on the surfaces of other cells (Garcia and Adams 2005). T lymphocytes have two major classes, differing by the presence of CD4 or CD8 co-receptor molecules on the T cell surface. The type of co-receptor molecule determines the class of MHC with which the corresponding T cell may interact. Cytotoxic CD8 T cells recognize peptides in the complex with MHC class I, which is an integral part of the surface of any cell type with minor exceptions. Such recognition results in removal of infected or mutated cells, presenting peptides of intracellular pathogens or of self-mutated proteins (Wong and Pamer 2003). CD4 T helper (Th) cells recognize peptides bound to MHC class II molecules, only occurring on the surface of specialized immune cells, which together with CD4 T helpers orchestrate the type and dynamics of immune response (Zhu, Yamane, and Paul 2010). It also includes activation of B cells followed by production of immunoglobulins (Crotty 2011).

B cell receptors (BCRs) recognize intact antigens and exist in two forms: in the form of membrane-bound receptor on the surface of B lymphocytes or in the form of soluble immunoglobulins or antibodies. The membrane-bound form is necessary for signal pathways and B cell fate decisions, such as differentiation in other B cell types or isotype switching. Soluble antibodies are involved in direct fight with pathogens: they neutralize antigens, preventing their functioning, form immune complexes (antigen-antibody complex) and invoke branches of innate immunity.

Both TCRs and BCRs have two chains in their structure ($\alpha\beta$ or $\gamma\delta$ chains in T cells and heavy and light chains in B cells), which include a constant region conservative among all lymphocytes of the same type and a variable lymphocyte-specific region (**Figure 2.1A**). In the case of B cells, the receptor is dimerised and consists of two heavy chains and two light chains of $\kappa$ or $\lambda$ type. Conservative regions cover part of the receptor, conducting the signal from the antigen binding site to the lymphocyte or other immune cell in the case of soluble immune complexes. Immunoglobulins have five main classes of constant regions called isotypes: IgM, IgD, IgG, IgA and IgE. They differ in distribution of immunoglobulin in the body and determine the reaction of other immune cells recognising it.

The variability of antigen receptors of T and B lymphocytes is not uniform along the length of the receptor amino-acid sequence (**Figure 2.1B**). Comparison of receptors reveals three hypervariable regions, called cluster determination regions (CDRs). In the folded structure of the receptor, both TCR and BCR, CDRs form hypervariable loops on the surface of the receptor right at the antigen-binding site (**Figure 2.1C**). Thus high diversity of CDRs covers a huge variety of antigens which immune receptors are able to specifically recognize. CDRs are separated by framework regions (FWRs), which being much more conservative than CDRs, still belong to the variable part of the receptor. FWRs play an important role in right orientation of CDR loops to each other and to the antigen (Zhou et al. 2020).

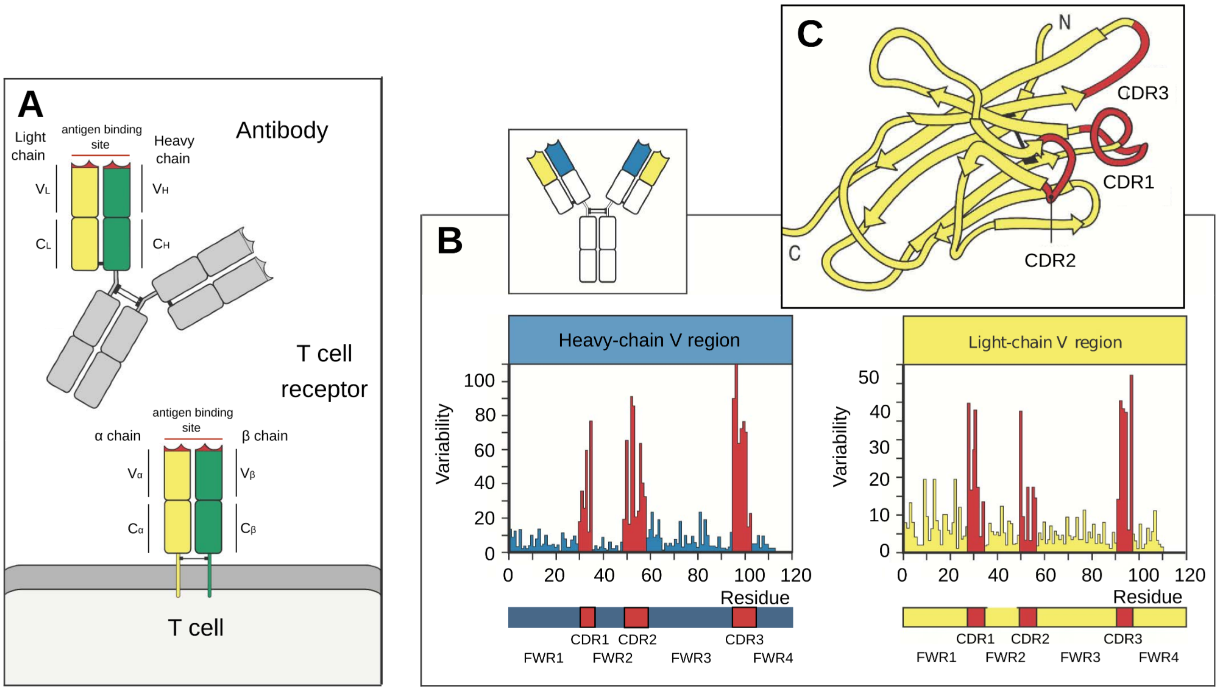


**Figure 2.1. Structure of lymphocyte antigen receptors.** **A**: Schematic structures of T and B cell receptors. Both receptors have two chains ($\alpha$ and $\beta$ or $\gamma$ and $\delta$ chains in T cells and heavy and light chains in B cells), which includes variable (V) and constant (C) regions; **B**: Distribution of variability, derived from comparison of amino-acid sequences of BCR variable regions. There are three hypervariable complementary-determining regions (CDRs), separated by more conservative framework regions (FWRs); **C**: Position of CDR regions in the 3D structure of the light immunoglobulin chain. They form hypervariable loops at the surface of the chain, where an antigen-binding site takes place. The figure is adapted from (Murphy and Weaver 2016).

### V(D)J recombination and diversity of antigen receptors

Such diversity of lymphocyte receptors is generated by the V(D)J recombination, followed by the lymphocyte clonal selection (Roth 2014; Chi, Li, and Qiu 2020; Jung and Alt 2004; Schatz and Ji 2011). Genes of TCR and BCR in immature lymphocytes are presented by the loci with sets of variable (V), diversity (D) and joining (J) segments (**Figure 2.2**). Loci of $\alpha$ chain of TCRs and heavy chain of BCRs include sets of segments of all three types. Loci of $\beta$ chain and light chains of BCRs are shorter and include V and J segments only.

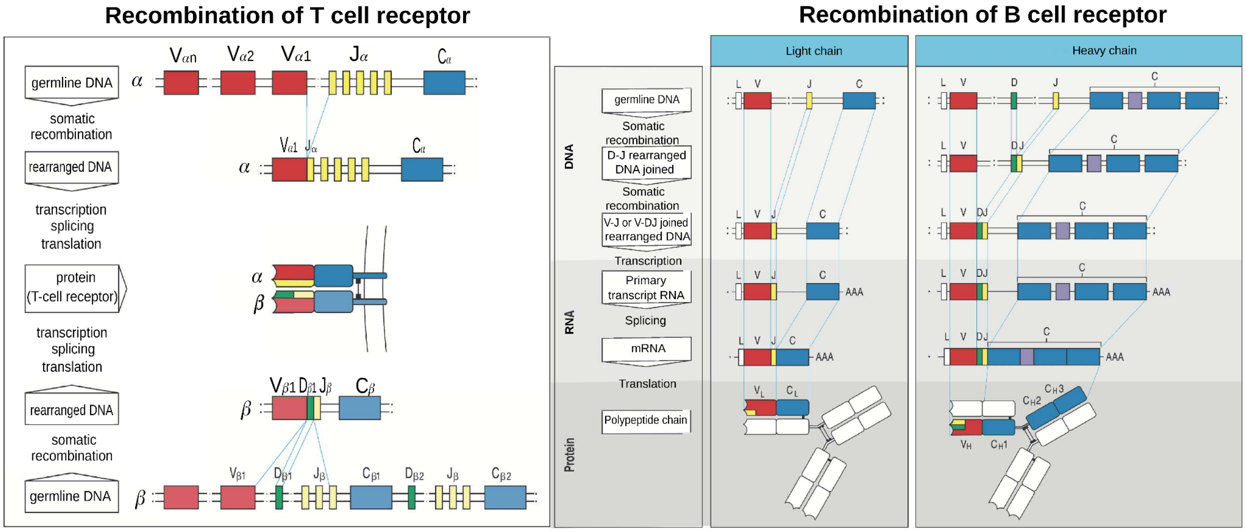


**Figure 2.2. Schematic representations of V(D)J recombination of T and B cell receptors.** The figure is adapted from (Murphy and Weaver 2016).

For a complete development of antigen receptors these loci should undergo gene rearrangement, resulting in a single segment of each of V, D and J types in TCR $\alpha$ (BCR heavy) chain and in a single segments of V and J types in TCR $\beta$ (BCR light) chain (**Figure 2.2**). Such precision in the set and order of segments in the resulting receptor is achieved by the 12/23 rule (Schatz and Ji 2011; Jung and Alt 2004). Each V, D and J segment in a loci is flanked by conserved heptamers, called recombination signal sequences (RSS). There are two types of RSS, with 23-base-pair and 12-base-pair spacers. Recombination and joining of segments normally occurs between segments, flanked by RSS of different lengths. The process is guided by the RAG 1/2 (recombination-activating gene) complex. It has two subunits with fidelity to different types of RSS. Thus RAG 1/2 initiates the process of segment joining, binding to RSSs of different lengths. The disposition of RSSs relative to VDJ segments in the receptor loci is organized in such a way, that subject to 12/23 rule there is no way to rearrange the receptor gene with a wrong segment composition. However, sometimes neighboring D segments can be recognized as a single one, resulting in a phenomenon of ultralong antibodies, generated by VDDJ combination of segments (Briney et al. 2019; Watson et al. 2006).

The part of diversity due to a random choice of segments in the process of VDJ rearrangement is attributed to combinatorial diversity and is estimated as $5.8 \times 10^6$ in receptors of $\alpha\beta$ T cells and $1.9 \times 10^6$ in immunoglobulins (Murphy and Weaver 2016). However there is another source of diversity, which also makes a significant contribution to overall diversity of antigen receptors. It is junctional diversity, which occurs at the borders of V(D)J segments. After the RAG 1/2 complex attaches to RSS sequences of two segments for following joining, it introduces double-stranded breaks in DNA between the segment and its RSS. Breaks then are closed in DNA hairpin ends. Next the complex of DNA-dependent protein kinase (DNA-PK) with Artemis enzyme opens these hairpins at random sites by single-stranded cuts, generating palindromic P nucleotides: a single-stranded tail from nucleotides of coding sequence followed by the complementary nucleotides from the opposite DNA strand. The DNA repair system, trying to restore DNA integrity, removes some nucleotides to pair single stranded tails together. At the same time terminal deoxynucleotidyl transferase (TdT), an enzyme specific to the lymphoid cell line, conversely adds random N nucleotides to the tails. As a result regions of nontemplate random PNP nucleotides between V-D and D-J segments in TCR $\alpha$ (BCR heavy) chain or V-J segments in TCR $\beta$ (BCR light) chain appear. They become the most variable parts of receptors and are located in the CDR3 region.

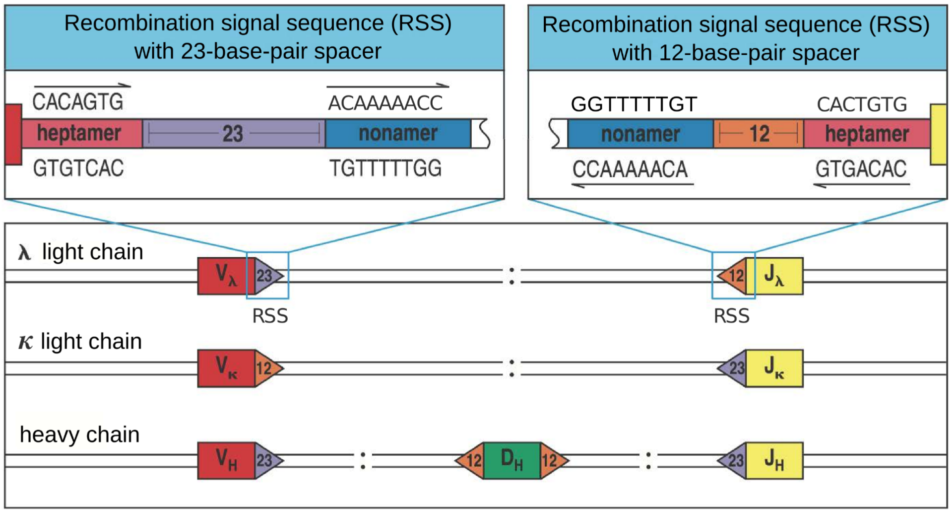


**Figure 2.3 VDJ segments are flanked by recombination-signal sequences (RSS), which guides the order of segments due to V(D)J rearrangement of the receptor gene.** The figure is adapted from (Murphy and Weaver 2016).

Naturally so many random events in receptor gene rearrangement may lead to the formation of unproductive gene sequences with frame shifts or inability to fold in the functional protein. In such cases lymphocytes undergo VDJ recombination one more time in the homologous chromosome. If the second try appears to be unproductive as well,  such lymphocytes go to apoptosis. DNA-sequences of BCRs revealed that most functional lymphocytes have an unproductive rearrangement on the second chromosome together with the functional BCR (Nourmohammad et al. 2019; Murphy and Weaver 2016). After VDJ rearrangement all lymphocytes go through clonal selection, where receptors with potential autoreactivity are also directed to apoptosis (Murphy and Weaver 2016; Roth 2014; Jung and Alt 2004).

### Affinity maturation of B cells

Affinity maturation (AM) is a part of the B cell immune response, which results in multifold improvement of specific B cell receptor (BCR) in its ability to recognize and bind the antigen (Teng and Papavasiliou 2007; Heesters et al. 2016; Chi, Li, and Qiu 2020). It occurs with T-dependent activation of B cells and is based on an evolutionary process consisting of repetitive cycles of somatic hypermutations (SHM) and clonal selection (Murphy and Weaver 2016; Kapila 2004).

B cell response starts from highly-specific recognition of the antigen by naive B cell, which triggers internalization of the antigen bound to the BCR. Next internalized antigen should be enzymatically processed and attached to MHC class II molecules for further antigen presentation (Heesters et al. 2016). Such B cells move in the lymph node to the border of B cell follicle and T cell zone, where they may interact with CD4 follicular helper T cells (Tfh cell), already differentiated in response to the same antigen (Crotty 2011). Specific recognition of MHC II:peptide complex by Tfh cell induces production of various cytokines in Tfh cell and expression of activatory ligands on its surface. It results in activation of the B cell, its further proliferation and differentiation. From this moment the BCR, involved in the primary recognition of the antigen, is present on membranes of numerous B cells, which altogether compose a B cell clone. Part of the clone forms a so-called primary focus. Some B cells move away from lymphoid follicles and differentiate into antibody-secreting cells (ASC): plasmablasts (PBL) or plasma cells (PL). Most B cells, involved in the primary focus, will not become long-living cells and eventually die. They are aimed to provide fast primary humoral response until high-affinity antibodies are formed.

At the same time another part of proliferating B cells stay at the lymphoid follicle together with associated Tfh cells and continue to proliferate, forming a germinal center (GC). GCs are the structures in which affinity maturation occurs. They develop during the first days of immune response and may still be present there in a month after the immune response starts. Activated B cells, involved in affinity maturation, repetitively migrate between two GC functional zones: dark and light ones (**Figure 2.4A**). In the dark zone a special enzyme, activation-induced cytidine deaminase (AID), introduces somatic hypermutations (SHMs) in the sequence of BCR genes (Chi, Li, and Qiu 2020). Deamination of cytosine leads to the formation of uridine in the DNA sequence, which results in the activation of various branches of DNA repair systems and substitution of C:G nucleotide pair (**Figure 2.4B**). By some estimates, 1% of SHM cases result in formation of indels (Teng and Papavasiliou 2007). Next the B cell with modified BCR moves to the light zone, where it can survive for a very limited amount of time: it should get a survival signal from the associated CD4 Tfh cell and come back to the dark zone or leave the germinal center. To do this, the B cell should bind and internalize the antigen, present in the limited amount on the surface of follicular dendritic cells (FDCs). Next, to get a survival signal B cell presents its epitopes in the complex with MHC II molecules to the corresponding Tfh cell. In the

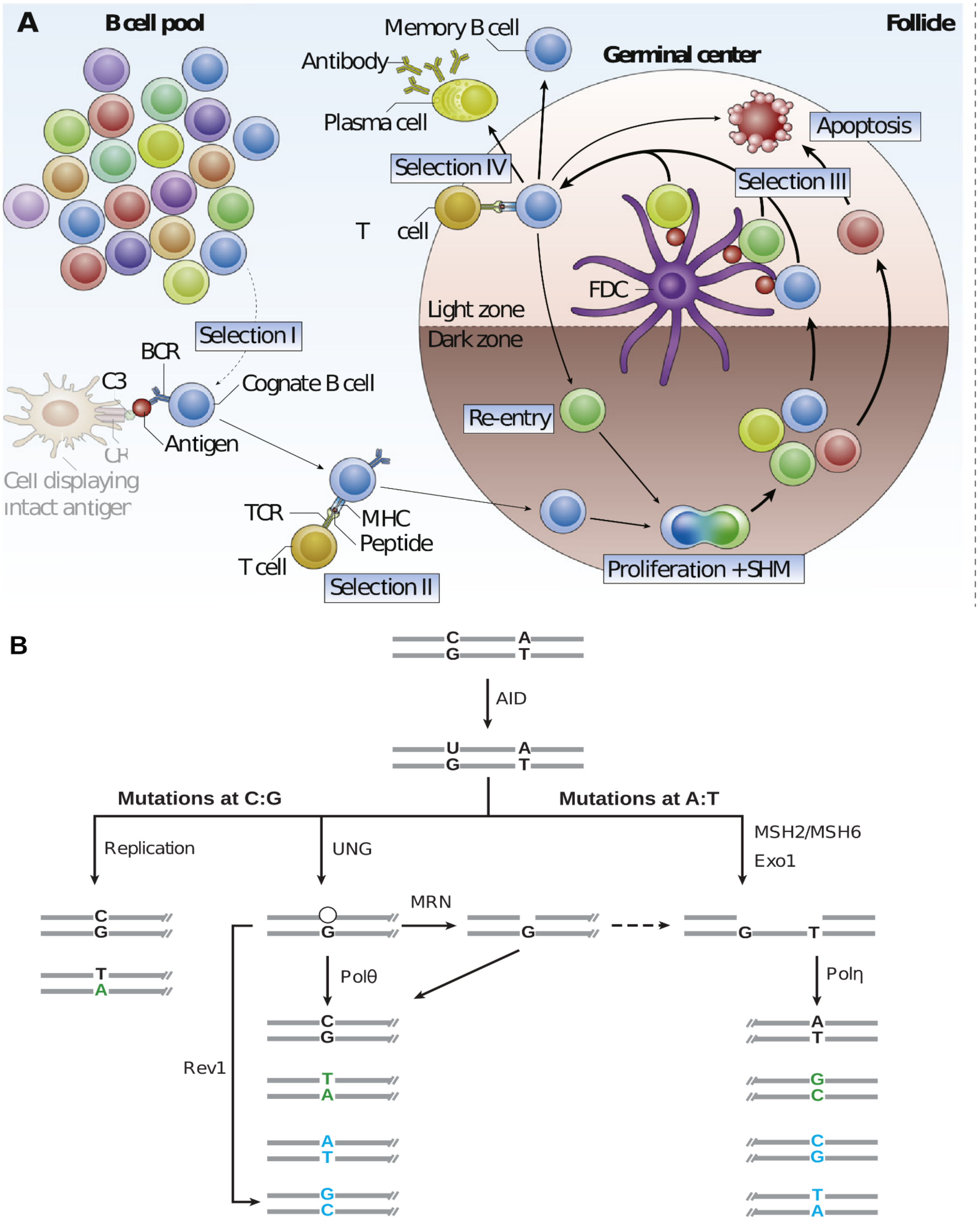


**Figure 2.4 Schematic representation of affinity maturation. A**: Model of affinity maturation in GC. Selection I - the step of initial recognition of the antigen by BCR with affinity, sufficient

for antigen binding and internalization; Selection II - B lymphocytes present processed antigens on MHC class II molecules on the T-B cell border of GC, get costimulatory signals and enter dark zone of GC for proliferation and somatic hypermutations; Selection III - BCRs with different SHMs compete for limited amount of antigen on the surface of FDCs; Selection IV - BCRs with the best affinity faster get survival signal from limited amount of Tfh cells and continue affinity maturation or leave the germinal centers. Cells that do not receive a survival signal undergo apoptosis; **B**: Deamination of cytosine by AID enzyme and induction of uracil in DNA sequence is followed by activation of different DNA-repair pathways, leading to transition (green) or transversion (blue) mutations. Panel A is adapted from (Heesters et al. 2016) and B from (Teng and Papavasiliou 2007).

case of negative effects of BCR changes on its binding affinity to the antigen, there will be no internalized epitopes to present and B-Tfh cross-recognition would not happen. In the absence of a survival signal B cells undergo apoptosis. Clonal selection intensifies as the number of involved B cells grows, since they start to compete with each other for the limited amount of the antigen and limited number of Tfh cells (Heesters et al. 2016). Therefore, if at the beginning of affinity maturation, clonal selection removes BCR variants that are unable to recognize the antigen, in later stages selection removes receptor variants that cannot bind antigen fast enough or have intermediate affinity. As a result of this process, a primary activated B cell gives rise to a B cell clonal lineage, composed of genetically close but different by accumulated SHMs B cell clones (Tas et al. 2016).

## Evolutionary analysis of B cell clonal lineages

### Introduction to immune repertoires

Immune repertoire is a collective diversity of B cell and T cell receptors in the organism, which is characterized by multifactorial and dynamic structure (Minervina, Pogorelyy, and Mamedov 2019). With development of high throughput sequencing it became possible to study immune repertoires on an unprecedented level (Liu and Wu 2018; Teraguchi et al. 2020). The group of techniques and its variations, used for sequencing of TCRs and BCRs, is called Rep-Seq (repertoire sequencing). Clonotypes refer to distinct B(T)CR sequences, which can exhibit significantly varied frequencies within the overall diversity of receptors. In case of BCR repertoires, clonotypes may be assigned to clonal groups or clonal lineages - sets of BCR sequences, originated from a single B cell and diversified due to accumulation of SHMs (Imkeller and Wardemann 2018).

Both genomic DNA (gDNA) and messenger RNA (mRNA) can be used for repertoire library preparation, however the choice depends on the purpose of the research (Liu and Wu 2018). gDNA libraries are easier to prepare and much better represent real proportions of lymphocytes, since the amount of gDNA per cell is not affected by varying expression levels of antigen receptors in different cellular subtypes. On the other hand, gDNA libraries include both productive and unproductive sequences of receptors, which can not always be distinguished by the nucleotide sequence only. gDNA libraries are usually based on the multiplex PCR amplification, using a set of specific primers for V or J genes (Klarenbeek et al. 2010). Different efficiencies and possible cross-reactivity of multiplex primers bias resulting libraries, which is especially the case for BCR repertoires because of SHM changes (Imkeller and Wardemann 2018).

Protocols for mRNA libraries possess more freedom to reduce the effect of primer bias. A widely used technique is 5′ rapid amplification of cDNA ends (5′RACE) (Scotto–Lavino, Du, and Frohman 2006). It starts from reverse transcription (RT) of mRNA with primers, specific for constant C-gene regions of T(B)CRs in the 3′ end of mRNA. In the end of RT additional nucleotides are added to the 3′ cDNA end through the template-switch process (Mamedov et al. 2013). In another variant of 5′RACE the RNA ligase adds a linker to the 3′ end of cDNA after RT. One way or another B(T)CRs can be amplified using C-gene and the universal linker primers (Gao and Wang 2015; Heather et al. 2016). Thus nowadays both multiplex PCR and 5′RACE techniques are widely adapted in the study of immune repertoires.

Furthermore such high diversity of B(T)CRs with drastically different concentrations of separate sequences greatly complicates detection of PCR and sequencing errors. This problem is especially important for the study of BCR repertoires, since variants created by affinity maturation are hard to distinguish from imperfections of sequencing protocols. To deal with this complication, a type of Rep-seq protocols with barcoding of clonotypes by unique molecular identifier (UMI) sequences were developed (Rosati et al. 2017; Shugay et al. 2014). In the example of MIGEC protocol, presented on the **Figure 2.5**, UMIs are introduced in cDNA sequences via the template switch at the end of RT (**Figure 2.5**). In further analysis, resulting sequences are distributed to molecular identifier groups (MIGs) - groups of sequences with the same UMI. Most PCR errors appear in late stages of amplification and such variants have low frequency in MIG. So the first stage of error correction is based on identification of the dominant sequence variant in each MIG  (**Figure 2.5B**). In the second stage of error correction those MIGs, presented by a single UMI, are filtered together with MIGs, where frequencies of two variants are comparable and the presence of dominant sequence is unclear.

Rep-Seq with specific amplification of antigen receptor sequences belong to the target technologies. Besides this immune repertoires can be extracted from the bulk RNA-Seq or Single-Cell sequencing data (Song et al. 2021, 4). Naturally the level of resolution of immune repertoires from bulk RNA is significantly lower than in case of target protocols, however with the amount and availability of open source RNA-seq data this opportunity may also be important for some research goals.

Single-Cell sequencing allows to track chain pairing in antigen receptors: heavy and light in BCRs and $\alpha$ and $\beta$ or $\gamma$ and $\delta$ in TCRs. It also allows to annotate clonotypes by phenotypes of T and B cells and to understand the type of immune response, in which particular clonotypes are involved. There are variations of Single-Cell protocols with additional enrichment of VDJ sequences of the data for immune profiling (Xu et al. 2022; R. D. Lee et al. 2021). Such an approach is now widely implemented in cancer research, where the structure and phenotype distribution of tumor infiltrating repertoires are promising predictive biomarkers for disease progression and efficiency of the therapy (Valpione et al. 2021).

Apart from specific features of sequencing procedure the post-analysis of immune repertoires may be also a challenging task. Clonotypes of repertoires are defined by the set of V(D)J segments, used in the gene rearrangement, as well as by nontemplate nucleotides, randomly inserted or deleted in junctions between them. Therefore determination of particular V(D)J segments, from which the clonotype was constructed, are particularly important for repertoire structure analysis (Odegard and Schatz 2006; Liu and Wu 2018). While the V and J segments are long enough to be definitely determined, the D segment may be just 12 nucleotides long (*Homo sapiens IGHD7-27*01*). Moreover, random events in junctions between segments due to VDJ rearrangement of $\alpha$ TCR or heavy BCR chains may lead to a complete loss of D specific features necessary for its recognition. In addition, VDJ genes are among the most variable genes in the human genome, and the diversity of VDJ alleles in the human population is dramatically poorly described (Mikocziova, Greiff, and Sollid 2021). Thus typically VDJ segments are determined to an accuracy of the segment gene family. In the case of BCRs, where clonotypes are additionally modified by SHMs, recognition of V(D)J segments with separation of SHM variants from germline VDJ alleles are even more challenging.

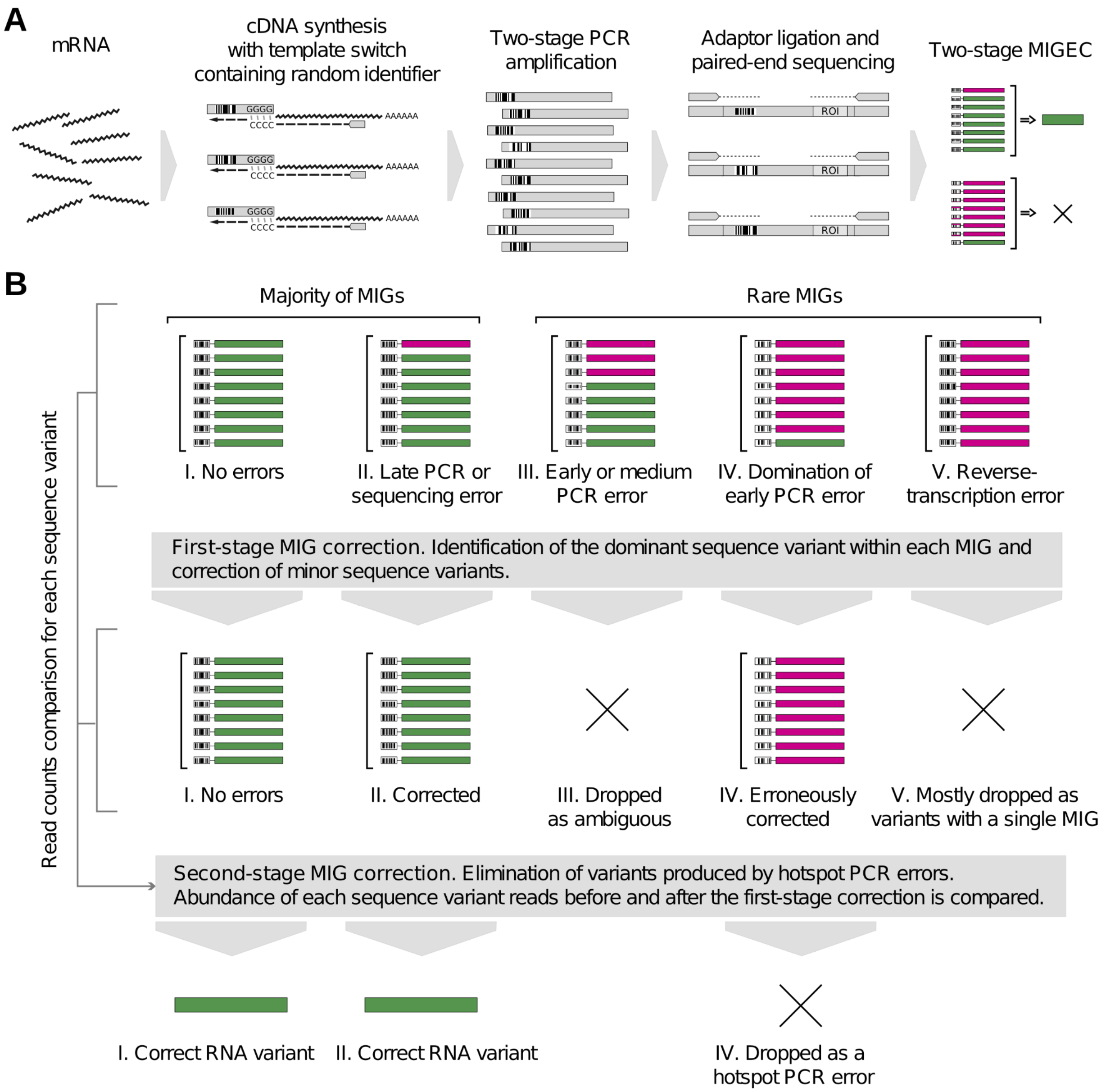

A
mRNA
cDNA synthesis with template switch containing random identifier
GGGG
CCCC
AAAAAA
Two-stage PCR amplification
Adaptor ligation and paired-end sequencing
ROI
Two-stage MIGEC
B
Majority of MIGs
Rare MIGs
I. No errors
II. Late PCR or sequencing error
III. Early or medium PCR error
IV. Domination of early PCR error
V. Reverse-transcription error
First-stage MIG correction. Identification of the dominant sequence variant within each MIG and correction of minor sequence variants.
I. No errors
II. Corrected
III. Dropped as ambiguous
IV. Erroneously corrected
V. Mostly dropped as variants with a single MIG
Read counts comparison for each sequence variant
Second-stage MIG correction. Elimination of variants produced by hotspot PCR errors. Abundance of each sequence variant reads before and after the first-stage correction is compared.
I. Correct RNA variant
II. Correct RNA variant
IV. Dropped as a hotspot PCR error

**Figure 2.5**. The scheme of the MIGEC pipeline (Shugay et al. 2014). **A**: The pipeline of UMI-based B(T)CR library preparation; **B**: The procedure of error correction in molecule identifier groups (MIGs).

Nowadays there are a bunch of bioinformatic tools, trying to solve these tasks. Some of them, such as MiXCR, pRESTO and TRUST4, are able to work with different types of sequencing pipelines, including bulk and target RNA-Seq, as well as Single-Cell protocols (Bolotin et al. 2015; Vander Heiden et al. 2014; Song et al. 2021). There are also special tools for identification of novel VDJ gene alleles, such as TigGER (Gadala-Maria et al. 2015) and IgDiscover (Corcoran et al. 2016). Immcantation framework, developed by AIRR (Adaptive Immune Receptor Repertoire) community, provides a whole ecosystem of tools for repertoire post processing. It includes tools of analysis of SHMs (SHazaM) and phylogenetic analysis of B cell lineages (dowser), physicochemical properties of antigen receptors (alakazam) and many others (Stern et al. 2014; Yaari et al. 2013; Gadala-Maria et al. 2015; 2015; Gupta et al. 2015). The Immunarch R package is also one of the useful repertoire post processing tools (Popov 2022). Thus the field of immune repertoires and community of scientists, involved in such types of studies, rapidly develop and greatly improve tools for the AIRR research.

**Peculiarities of B cell clonal phylogeny**

Immune repertoires of antigen-experienced B cells are structured in lineages, originating from a single naive B cell. Lineages diversify due to the process of affinity maturation after a single or several antigen challenges (Tas et al. 2016). In their recent study Yermanos and colleagues introduced the term antibody forest, reflecting this structure (Yermanos et al. 2020). Indeed each B cell clonal lineage has the history of its evolution, which is tracked on the phylogenetic tree. Thus phylogenetic trees of all lineages in a repertoire compose a forest. Common evolutionary biology tools and pipelines can be applied to antibody forests as well, however some specific features of B cell clonal evolution should be taken into account to design the analysis.

The first peculiarity is that mutation rate in B cell clonal lineages (in fact the rate of SHM) is relatively high and according to some estimates riches $10^{-3}$ mutations per base per cell division. Such value exceeds the spontaneous mutation rate in somatic cells by a factor of $10^{6}$ (Odegard and Schatz 2006). The only organisms with comparable mutation rates are RNA viruses, which can change with $10^{-6}$ - $10^{-4}$ substitutions per nucleotide per cell infection (Duffy 2018; Peck and Lauring 2018; Drake and Holland 1999). The distribution of variability in BCRs is not uniform and most of the density of SHMs falls into the CDR3 region. CDR3 also possesses high junctional diversity, since it includes nontemplate nucleotides between V(D)J segments. These two facts together complicates the procedure of clonal lineage assignment: clonotypes, which differ because of two independent recombination events from the same V(D)J segments, should be distinguished from SHM differences (Odegard and Schatz 2006).

To deal with the problem of clonal group assignment, there is an approach of clonal group similarity threshold determination, realized in the ChangeO toolkit of the Immcantation framework (Gupta et al. 2015). The approach is based on the distribution of DNN (distances to the nearest neighbor) in the repertoire. The nearest neighbor in the repertoire is the clonotype with the minimal number of nucleotide differences from the set of clonotypes with the same VDJ segments and the same CDR3 length. The distribution of distances to such neighbors appears to be bimodal: in average clonotypes of the same lineage differing by SHMs are less distant from each other than clonotypes, originating from very similar but independent VDJ rearrangements (**Figure 2.6**). Thus the border between these two modes can be used as the maximum threshold of the clonal group similarity. The Change-O tool may use different distance models depending on the needs of the user.

The procedure of clonal group assignment usually is based on the following criteria: 1. clonotypes of the same lineage should be rearranged from the same VDJ segments; 2. should have CDR3s of the same length; 3. should satisfy a determined similarity threshold. Similarity threshold varies between studies, however it usually falls in the range of 80-90% (Hoehn et al. 2021; Horns et al. 2019; Nourmohammad et al. 2019). The recent study revealed that moderate changes in lineage criteria affect the distribution of sizes of antibody trees, but does not influence biological conclusions from the data analysis (Yermanos et al. 2020).

Another specific feature of B cell phylogenies is that antibody trees can be rooted by the germline sequence of the corresponding VDJ segments (Odegard and Schatz 2006; Barak et al. 2008). The primary BCR sequence, from which the lineage evolution starts, can not be fully determined, because of nontemplate nucleotides in the junctions between V(D)J segments. Nevertheless, even in the case of short BCR clonotypes the germline V(D)J positions comprise a significant part of the clonotype and may define the direction of clonal evolution on the phylogeny (Yermanos et al. 2020). Rooting of B cell phylogenies by V(D)J germline constructs allows to track the order of SHM occurrence and branching events in the lineage, so it is widely used in clonal lineage analysis (Nourmohammad et al. 2019; Hoehn et al. 2021; Horns et al. 2019; Yermanos et al. 2020; Kräutler et al. 2020; Yaari et al. 2013).

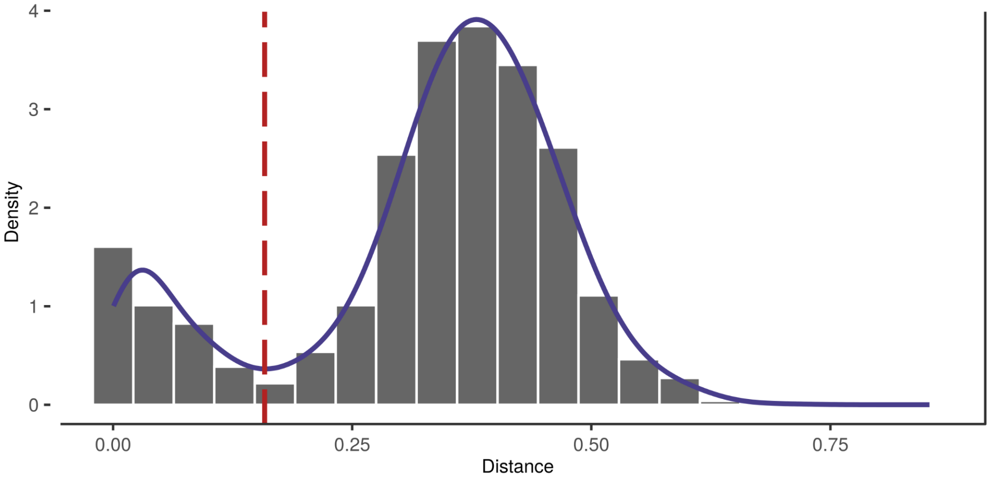


**Figure 2.6**. The bimodal distribution of distances to the nearest neighbor in a repertoire, which allows to determine the threshold of clonal group similarity. The figure is adapted from (Gupta et al. 2015).

Analysis of longitudinal repertoire data revealed that the dynamics of SHM accumulation is not linear and rarely corresponds to the timeline of clonotype sampling dates. Indeed affinity maturation and SHM accumulation do not proceed permanently with the constant rate and depend on antigen challenges. Following this logic Hoehn and colleagues developed a criteria to distinguish vaccine-responding B cell lineages from vaccine-independent: the timeline of clonotype sampling in vaccine-responding B cell lineages is correlated with the number of accumulated SHMs (Hoehn et al. 2021). Thus early clonotypes are located closer to the root of the tree, when clonotypes from later time points tend to sit on longer branches (**Figure 2.7A:B**). In vaccine-independent lineages the time point of the clonotype sequence is unrelated to the clonotype position on the lineage phylogeny (**Figure 2.7C:D**).

The feature of B cell clonal lineages to include both terminal and internal node sequences of phylogenetic tree is often noted (Davidsen and Matsen 2018; Barak et al. 2008; Odegard and

Schatz 2006). Indeed not all B cells in the lineages accumulate SHMs with the same rate and after cellular division one of daughter B cells may ‘fix’ the sequence of the BCR and proliferate further without introduction of new changes in its antigen receptor. At the same time progenitors of its sister cell will still diversify due to affinity maturation. Thus in the sample of BCRs from a single lineage, one may observe both ancestral and descendant BCR sequences of lineage phylogeny.

Another feature, which is also frequently mentioned in the literature, is the mutational signature of the AID enzyme, which produces the diversity of BCRs for selection due to affinity maturation. It defines the probability distribution of SHMs to occur in different nucleotide contexts (Teng and Papavasiliou 2007; Pettersen et al. 2015; Rogozin et al. 2016).

Fortunately, evolution of B cells is not unique within the context of these factors. When sampling viral populations, some sequences also appear in the internal nodes of the tree due to interrupted epidemiological chains (Komissarov et al. 2021; Klink et al. 2021; Hall, Woolhouse, and Rambaut 2016). Specific mutational signatures and the role of certain enzymes in mutagenesis and in population evolution is well described for viruses and tumor cells (Graudenzi et al. 2021; Yi et al. 2021) and tumors (Y.-A. Kim et al. 2021; Koh et al. 2021). Modern phylogenetic tools, based on maximum likelihood or bayesian approach, take into account both varying mutation rate on different phylogenetic branches and population-specific patterns of mutagenesis (Drummond et al. 2006; Yang 2006; Nei and Kumar 2000).

One more specific feature of B cell phylogeny is that BCR diversification simultaneously occurs on both productive and unproductive rearrangements. Unproductive V(D)J rearrangements accumulate SHMs together with their homologous BCRs with the only difference - SHMs in unproductive sequences have no effect on B cell fitness and fate. Thus they accumulate neutrally at the rate of the mutational process itself and hitchhike together with its productive neighbor. Therefore in the case of gDNA based Rep-Seq unproductive BCRs may be used for estimation of SHM rate or for estimation of the dN/dS values  (the ratio of nonsynonymous to synonymous changes) in sequence evolution in the absence of selection. Several studies used unproductive sequences as a neutral baseline for productive BCRs, evolving under action of natural selection (McCoy et al. 2015; Nourmohammad et al. 2019).

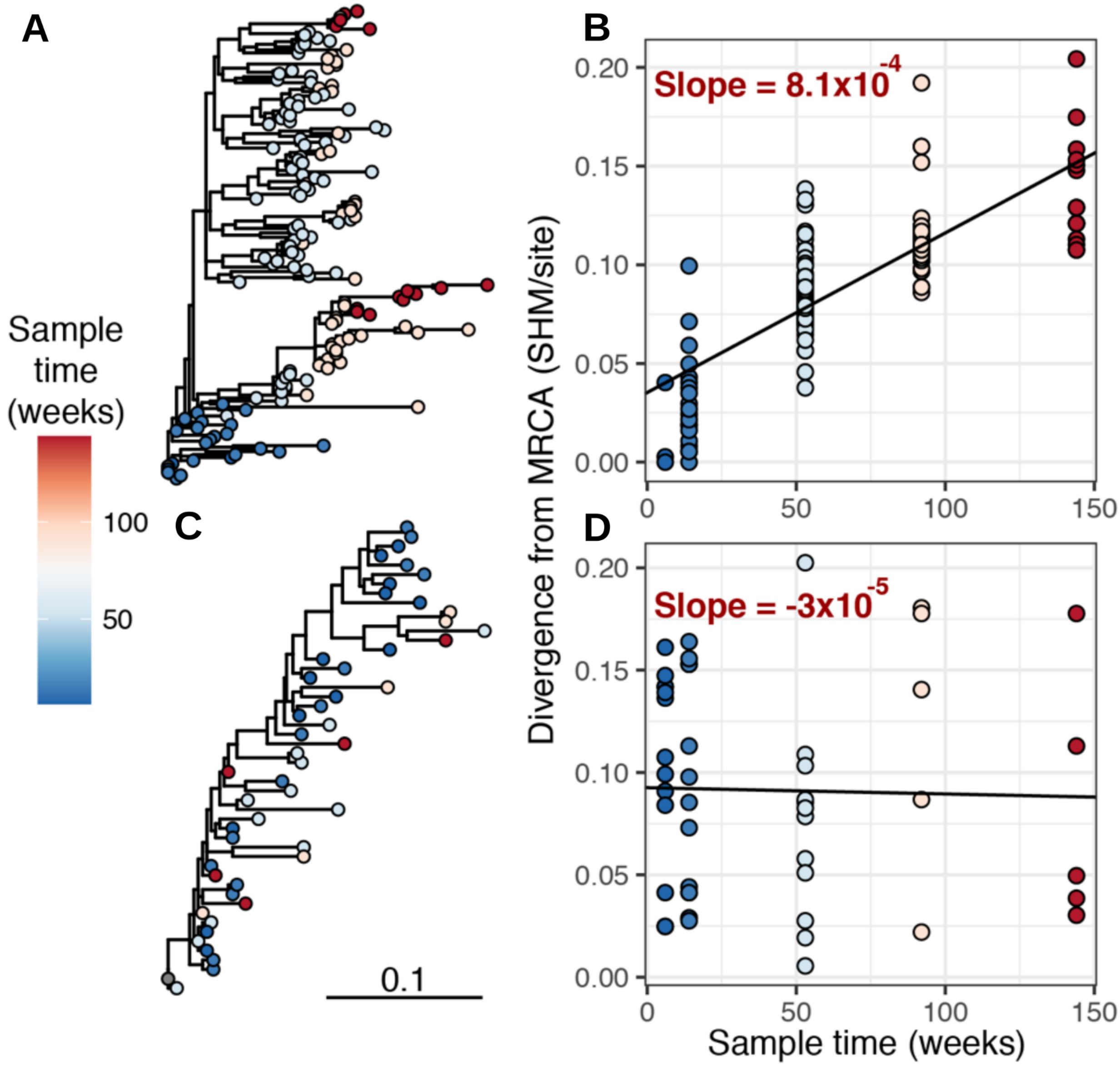


**Figure 2.7** Example of vaccine-responding (**A**) and vaccine-independent (**C**) B cell lineage phylogenies. The color of leaves on the tree corresponds to the sampling time of the clonotype. The gray dot corresponds to the germline root of the B cell tree; **B** and **D**: Correlation between the rate of accumulated SHMs in clonotype and its sampling time of trees from A and D panels; The figure is adapted from (Hoehn et al. 2021).

### Studies of B cell clonal evolution

In an increasing number of works, BCR repertoires are studied with the use of techniques of populational genetics. Commonly such studies are focused not on a whole BCR repertoires, but on the most abundant B cell lineages. Here are several examples of how evolutionary analysis of B cell lineages may be used to address questions about the work of B cell adaptive immunity. Horn and colleagues observed B cell clonal lineages after influenza vaccination (Horns et al. 2019). In their work, peripheral blood of 5 healthy adults was sampled nine times from the fifth day before the vaccine induction to the eleventh day after. Among the most abundant clonal lineages, authors distinguished persistent and vaccine-responsive ones by the dynamics of their fraction in repertoire through time. In persistent lineages there was a predominance of clonotypes with IgM or IgD isotypes, when vaccine-responsive lineages were mostly composed of switched IgA or IgG ones. Using models of Kingman (Kingman 1982) and Bolthausen-Sznitman coalescence (Bolthausen and Sznitman 1998; Neher, Kessinger, and Shraiman 2013) they simulated scenarios of neutral evolution, neutral evolution with expanding population and adaptive evolution for populations of sizes, same to sizes of B cell lineages. They revealed that distributions of SHMs in persistent lineages are well described by a neutral model, while vaccine-responsive lineages fit well the dynamics predicted by the model of adaptive evolution.

Furthermore, the authors have shown that vaccine-responsive lineages expanded more on the seventh day after vaccination. They also revealed affinity-enhancing and affinity-diminishing mutations, analyzing changes in branching rate on B cell phylogenetic trees and using the principle described in (Neher, Russell, and Shraiman 2014). An excess of affinity-enhancing mutations occurred in the CDR3 region of the BCR, when affinity-diminishing mutations were mostly revealed in the CDR1 and CDR2 regions.

The problem of finding BCRs with the best affinity in the lineage was raised in another work of Ralph and Matsen (Ralph and Matsen 2020). The study is based on a dataset with measured neutralization ability of different BCR variants of the same lineage. Antibody neutralization activity is well correlated with its affinity and may be used as a quantitative representative of the BCR fitness. Authors tested several phylogenetic metrics, as a predictor of the BCR fitness, including the Hamming distance of the sequence from the germline, the local branching index on phylogenetic tree and the number of accumulated SHMs. The local branching index together with the amino-acid distance of the sequence from the lineage consensus have shown best results. Indeed consensus of the lineage is close to the sequence with the maximum local branching index, so these two metrics are meaningfully close. These observations are well consistent with the previous article of Horns et al. and pipelines developed for fitness prediction in viral populations (Neher, Russell, and Shraiman 2014).

In another study of Nourmohammad et al. there was a longitudinal analysis of B cell clonal lineages of HIV infected individuals without treatment and with interrupted ART-therapy (Nourmohammad et al. 2019). Here BCR repertoires were sequenced from gDNA and authors used lineages of unproductive BCR sequences as a baseline for SHM rate. BCR lineages of HIV infected individuals also carried the signature of positive selection, which was correlated with the viral load in patients with interrupted ART-therapy. In addition, authors developed a beautiful pipeline to analyze effects of clonal interference in longitudinal BCR repertoires, based on likelihood dN/dS ratios for mutations with different frequency dynamics. The effect of clonal interference was pronounced the most in CDR3 region and slowed down the fixation of the most beneficial BCR variants.

Phylogenetic tools can also be used for reconstruction of the course of events in the lineage, such as isotype switching. In his other work, Horns and colleagues have shown that B cells with close BCR sequences tend to switch in the same clonotype, however they lose coherence in isotype switching with BCR diversification (Horns et al. 2016). Thus in general the process of isotype switching is B cell specific and is not defined uniformly for the whole lineage. Also the feature of isotype to switch in a particular order in most cases should be reflected in the order of isotypes on lineage phylogeny and can be used as the quality control of reconstructed phylogeny (Davidsen and Matsen 2018).

In addition to isotypes, B cell lineages can be annotated by the tissue of sampling or by a particular cell type, from which clonotype was obtained. Hoehn and colleagues have shown memory reactivation of B cell lineages due to influenza revaccination: clonotypes of resting memory gave rise to a clade of clonotypes from the germinal center (Hoehn et al. 2021). Such observations lead to the conclusion that a new antigen challenge is able to force B cell immune memory to start new cycles of affinity maturation and readapt to evolved antigen.

Tracking of evolution of B cell clonal lineages is also a promising approach in the investigation of B cell lymphomas (Küppers 2005). The group of Rachael J.M. Bashford-Rogers published several studies, devoted to the analysis of B cell repertoires of patients with chronic lymphocytic leukemia (CLL) and follicular lymphoma (FL) (Bashford-Rogers et al. 2013; Petrova et al. 2018). They have shown that malignant lineage usually arises from a single B cell clonotype, which remains the dominant clonotype with the disease progression. However it diversifies with time, producing numbers of subclones and forming the malignant lineage, occupying almost a whole B cell repertoire of the patient (Bashford-Rogers et al. 2013). They also detected the case of the CLL patient with two independent malignant lineages.

CLL malignant B cell lineages strongly differ from both B cell lineages of healthy individuals and other non-malignant lineages of same patients.They were predominantly composed of IgM isotypes, which is unusual for diversified lineages in healthy repertoires. They also possessed an unprecedentedly high number of accumulated SHMs in comparison with IgM clonotypes of healthy repertoires (Petrova et al. 2018). The landscape of isotype switching in malignant B cell lineages dramatically differs from healthy lineages as well: the switch from IgM/D to IgA1/2 predominates in malignant lineages, when in healthy B cell clonal lineages the switch frequency from IgM/D is almost equally distributed between IgA1/2 and IgG1/2. However authors observed no association between frequency of isotype switching in the malignant lineage and its relative size in B cell repertoire.

## Viral escape from the adaptive immunity

### Viral ways to escape adaptive immune response

The development of such a complex and beautiful system of adaptive immunity would be unnecessary without the adaptive evolution on the other side of the host-pathogen universe. Throughout their whole history, pathogens elaborate new tricks to avoid mechanisms of host immune defenses. Such an arm race spurs the rate of adaptation in both pathogens and the host, so in some cases it becomes possible to observe host-pathogen coevolution in a real time. In this context viruses, as the most fast-evolving organisms, are of particular interest.

Viruses are a huge and diverse group of organisms, which use replication systems of host living cells for the production of new viral particles. Viral escapology includes a whole set of strategies to deceive and avoid host defenses. Roughly these strategies may be grouped in a ‘camouflage and sabotage’ and ‘speed and shape-change’ principles (Lucas et al. 2008). The ‘camouflage and sabotage’ strategy is more common in DNA-based viruses with large genomes, which possess enough space to encode highly-evolved molecules allowing viruses to persist in the host imperceptibly for the immune system. Herpesviruses are a good example of ‘camouflage’ principle: their double-stranded DNA genomes may include over 200 viral proteins and have slow rates of replication. However many of these proteins are aimed at hiding from the immune system by latency in the form of episomal DNA (Connolly, Jardetzky, and Longnecker 2021). Thus after primary infection such viruses may persist for decades with no special damage to the host organism.

‘Speed and shape-change’ are usually short RNA-based viruses. They have no capacity to develop complex hiding mechanisms, however high rates of their replication allow them to rapidly modify their antigens, which target adaptive immune response (Lucas et al. 2008). Such modifications may include both changes of antibody binding sites to escape B cell immunity or changes of viral epitopes, presented on MHC molecules for T cells. In the case of B cell escape only structural and envelope proteins can be involved. Escape from neutralizing antibodies was described for many fast-evolving viruses including HIV-1 (Meijers et al. 2021; X. Wei et al. 2003; Dingens et al. 2019), Hepatitis B and C viruses (Lazarevic et al. 2019; von Hahn et al. 2007), influenza (Krammer 2019; Gentles et al. 2020; Leon et al. 2017), SARS-CoV-2 (Chakraborty et al. 2022; Hu et al. 2022; Weisblum et al. 2020; Harvey et al. 2021), LCMV (Eschli et al. 2007; Ciurea et al. 2001) and many others. In cases of chronic viral infections such host-pathogen coevolution may be tracked inside a single host: viral immunoediting of antibody-binding sites is followed by affinity maturation of corresponding B cell lineages with the adaptation of BCRs to changes in the antigen (Bonsignori et al. 2017; Muecksch et al. 2021). In cases of short-term viral infections with high transmissivity such as flu or COVID-19 some of antibody escape variants turn out to be universal, recurrently appear in various hosts and rapidly spread in a host population. Dynamics of such variants are tracked with an unprecedented precision during the current pandemic of SARS-CoV-2 (WHO 2022).

Escape from the T cell adaptive immunity implies prevention of recognition of the MHC:epitope complex by activated T cell clones, which can be done on different stages of presentation of viral peptides. First, viruses may change proteasomal cleavage sites used for intracellular generation of viral epitopes during antigen processing (Allen et al. 2004; Draenert et al. 2004; Kimura et al. 2005). However this strategy is not reliable, since peptides may naturally degrade and the epitope of concern may time to time appear in the intercellular matrix. Next, viruses can modify anchor positions in peptide antigens, which is used for binding to MHC molecules. In such a case activated T cell clones would be unable to recognize corresponding antigens because of their absence on a surface of infected cells. And the last escape modification may affect epitope immunogenicity - ability to be recognized by the specific T cell receptor. According to some estimates TCRs are able to recognize just a half of viral epitopes, presented on MHC complexes. Bronke and colleagues have shown that HIV effectively uses all these strategies, however prevention of epitope binding to MHC molecules is the most reliable escape mechanism (Bronke et al. 2013). In addition to viruses the same tactics are used by tumor cells (Marty et al. 2017).

Since escape ways from T cell clones are highly dependent on the set of HLA alleles of a particular individual, the existence of universal T cell escape viral variants is under question. However the distribution of HLA allele frequencies is not uniform and usually every human population has highly frequent HLA alleles, which would be present in almost a half of individuals (Buhler and Sanchez-Mazas 2011; Maróstica et al. 2022). Therefore escape from the most frequent HLA alleles can be considered as universal to some extent.

### SARS-CoV-2, a novel coronavirus

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first detected in December 2019 and was declared as pandemic on March 11, 2020 (Halaji et al. 2021; Mendiola-Pastrana et al. 2022). SARS-CoV-2 belongs to coronaviruses, the highly diverse family of enveloped positive single-stranded RNA viruses, infecting mammals and avian species. By the beginning of the pandemic, there were eight strains of coronaviruses known to infect humans, two of which, SARS-CoV in 2002 and MERS-CoV (Middle East respiratory syndrome coronavirus) in 2012, had caused severe outbreaks (Feng et al. 2009; Zumla, Hui, and Perlman 2015). A SARS-CoV-2 infection results in a coronavirus disease 2019 (COVID-19), which in most cases flows asymptomatically or as mild respiratory pathology; however, in a minor fraction of cases, the immune response to SARS-CoV-2 provokes hyperinflammation, leading to systemic multi-organ collapse (Mendiola-Pastrana et al. 2022).

The SARS-CoV-2 genome is about 30 kb long and encodes 29 genes. SARS-CoV-2 virion consists of four structural proteins (N, E, M and S). The positively-charged N (nucleocapsid) protein is responsible for a tight packaging of the viral RNA genome inside the nucleocapsid structure. E (envelope) and M (membrane) proteins together with the S (spike) protein incorporate in a lipid bilayer and form a viral envelope. The S proteins form trimmers, which stick out the membrane as corona-like “spikes”. SARS-CoV-2 enters human cells using the Receptor Binding Domain (RBD) of these outside parts of the S proteins through the angiotensin-converting enzyme 2 (ACE2) on the host cell surfaces. After entry to the host cell, two large open reading frames (ORF1a and ORF1b) are immediately translated into two polyproteins, which are then post-translationally processed in a set of individual nonstructural (nsp) proteins. Nsp proteins are in charge of formation of the viral replication and transcription complex and include RNA dependent RNA polymerase and exonuclease, responsible for high-fidelity replication (Robson et al. 2020).

During the SARS-CoV-2 pandemic, the scientific community has detected an enormous number of viral mutations (Shu and McCauley 2017; Knorre et al. 2021; Gangavarapu et al. 2022; Tsueng et al. 2022; Hodcroft Emma 2020). On average each nucleotide position of SARS-CoV-2 genome has at least one mutation (J. Chen, Wang, and Wei 2021). Such variability in SARS-CoV-2 population arises from a large number of factors, such as infidelity of replication and transcription of the viral genome (V’kovski et al. 2021; J. Chen, Wang, and Wei 2021), viral and host-viral recombination (Turakhia et al. 2022; Wertheim et al. 2022), host editing (Gribble et al. 2021) and so on. However, the nsp14 enzyme together with RNA-dependent RNA polymerase (nsp12) possess a proofreading activity, which distinguishes SARS-CoV-2 from other RNA viruses, such as flu and HIV, making its mutation rate somewhat slower ($1\times10^{-3}$ substitution/site/year) (Astuti and Ysrafil 2020; Koyama, Platt, and Parida 2020). Evolutionary forces, including random genetic drift, gene flow and natural selection (J. Chen, Wang, and Wei 2021) together with epidemiological factors, influencing dynamics of viral population size and ways of variant migration (Komissarov et al. 2021; Klink et al. 2021), shapes dynamics of this diversity. Thus, many variants demonstrated rapid global or local increase in their frequency in a viral population, suggesting probable beneficial effects on the fitness of the virus. For example, the famous S:D614G mutation was detected at early 2020 for the first time and was already present in all viral sequences until June of 2020. On the contrary, some variants were replaced by other variants shortly after their emergence. However their actual effect on viral phenotype and key factors, affecting their dynamics, is not always clear.

SARS-CoV-2 variants, which possess mutations with known effect according to the World Health Organization, such as increase of viral transmissivity, immune or diagnostic escape or association with the higher risk of COVID-19 severity belong to the group of VOI (Variants of Interest) (Hodcroft Emma 2020; WHO 2022). VOI, which became responsible for new COVID-19 waves, and in some cases globally displaced all other SARS-CoV-2 variants, are called VOCs (Variants of Concern). During the SARS-CoV-2 pandemics there were several waves, named after letters of the Greek alphabet: Alpha, Beta, Gamma, Delta, Theta, Epsilon, Kappa, Lambda, Mu and Omicron (Dutta 2022). Interestingly, many mutations characterizing these VOCs are recurrent and arose independently in different VOC lineages. For example, modification of residues 452 or 501 of the RBD domain in Spike protein, resulting in significant increase of viral transmissivity (J. Chen et al. 2020), is present in most SARS-CoV-2 lineages. S:L452 is present in Lambda (S:L452Q) and Delta (S:L452R), while S:N501Y is present simultaneously in Alpha, Beta, Gamma and Omicron (Hodcroft Emma 2020).

The first SARS-CoV-2 variant, which became the first globally circulating strain, differed from the Wuhan virus by a single S:D614G amino-acid substitution (Korber et al. 2020). S:D614G independently occurred in different parts of the world and spread globally. It is located outside the RBD region and improves viral replication (Plante et al. 2021). The next Alpha variant (B.1.1.7) emerged in the United Kingdom in September 2020 and already accumulated 23 mutations, including S:D614G. Among other notable mutations there were two deletions in the S protein: S:Δ69-70HV and S:del144_144. Both of them are highly recurrent and were detected in other VOCs and individual patients. S:Δ69-70HV is also one of mutations from the ΔF combination, associated with the outbreak in mink farms (Oude Munnink et al. 2021). In general, the Alpha variant was characterized by higher transmissivity and higher risk of severe COVID-19 than the original virus (Lyngse et al. 2021). At the same time there was an outbreak of Beta variant (B.1.351) in South Africa, detected in May 2020 for the first time. The next Gamma variant (P.1) emerged in Brazil with the first documented sample, dated by November 2020. Both Beta and Gamma variants demonstrated reduction of neutralizing activity of mAb and convalescent plasma (X. Chen et al. 2021; Souza et al. 2021), however their outbreaks mostly affected regions of their emergence and were not global.

In contrast, the next Delta variant (B.1.617.2), while started in India in October 2020, rapidly became a predominant variant in most countries. It includes 32 mutations, 12 of which are in the Spike. S:L452R and S:T478K mutations in RBD region with T19R, G142D, D614G, P681R, 157del, R158G, 156del, and D950N were shown to increase Delta transmissivity in two times in comparison with Alpha variant (Chakraborty et al. 2022). Some of them, such as S:L452R, have a pleiotropic effect on viral phenotype, including both increase of viral infectivity and escape from neutralizing antibodies (Deng et al. 2021). Specific feature of the Delta variant is its high disease severity, caused by more stable interaction of the RBD domain with ACE2, resulting in more effective entry into lung cells. The Delta variant has a 108% higher risk of hospitalization than the original virus and a 133% higher risk of mortality (Shiehzadegan et al. 2021).

Later, in November 2021, another highly divergent Omicron variant (B.1.1.529) was detected in South Africa, which became the most fast-spreading virus ever discovered (S. Kim et al. 2021). Omicron's genome accumulated 50 mutations with the most part of them (32) in S protein. Many of these mutations match with mutations of previous SARS-CoV-2 variants, suggesting that Omicron grouped properties of previous waves all together. Several studies reported that S:N679K, S:N501Y, S:P681H, S:N679K, and S:D614G changes in RBD region increase viral transmissibility, which exceeds transmissibility of the most aggressive previous Delta variant by a factor of 2.5-3.5 times (Gong et al. 2021). Deletions in ORF1a protein (L3674-, S3675-, and G3676) were shown to prevent destruction of viral components by intercellular innate immunity (Walls et al. 2020). Moreover, the Omicron variant possesses a much higher replication rate; however replication of the Omicron variant is more effective in the upper respiratory tract and less effective in lung tissues, which decreases the risk of COVID-19 severity (Mohapatra et al. 2022).

## Intrahost evolution of SARS-CoV-2

A whole set of studies are devoted to the effects of mutations of major SARS-CoV-2 VOCs on viral fitness, which in general results in improvement of viral transmissivity and replication rates, escape from innate immunity and modification of binding sites of neutralizing antibodies (J.

Chen et al. 2020; Gong et al. 2021). Many of these mutations, possessing beneficial effects on viral fitness and being positively selected, occurred independently in several SARS-CoV-2 strains and in individual patients with long-term COVID-19 (**Figure 2.1**). Intuitively, one would expect that at least a part of such mutations should be fixed gradually in the viral population, as it was with the first globally circulating S:D614G variant. However, most SARS-CoV-2 strains rapidly arised from a highly diverged variant, different from all other known variants by its own specific set of accumulated mutations. Such dynamics of SARS-CoV-2 evolution gives the reason to hypothesize that major SARS-CoV-2 strains accumulated their mutations in individual patients and not in the global population (Kupferschmidt 2021). In this context, the cases of long-term COVID-19, when viral evolution inside a single host can be tracked for some period of time, become of particular interest (**Table 3.1**).

Typically, viral RNA can be extracted from airway fluids, blood and feces between 3 and 46 days after symptom onset (Fu et al. 2020; Qian et al. 2020), however usually on 8th day infectious SARS-CoV-2 particles are already absent (Wölfel et al. 2020). The condition of long-term COVID-19 is characterized by a long period of not just viral RNA shedding, resulting in positive PCR tests, but also by the presence of infectious viral particles. In some cases such a period of viral persistence can take several months or even a year (Williamson et al. 2021; Monrad et al. 2021; Cunha et al. 2021; Borges et al. 2021) and can be separated by intervals of negative PCR tests as in case from (Sepulcri et al. 2021). In some cases such a long period of viral persistence is caused by the reinfection on the background of primary disease (Tillett et al. 2021; Prado-Vivar et al. 2021; Mulder et al. 2021). However, in most cases the track of viral evolution confirms that viral particles, detected after negative PCR test periods, belong to the same evolutionary track of the original infection (Avanzato et al. 2020; Kemp et al. 2020; Cele et al. 2021).

During the pandemic, a whole set of case studies, describing long-term COVID-19 has been reported (**Table 3.1**). Such cases are reported for individuals with various immune conditions. Long-term COVID-19 may occur in people with usual immune status and may last up to 112 days (Agarwal et al. 2020; Tillett et al. 2021; Prado-Vivar et al. 2021; Voloch et al. 2021). However, the most impressive duration of viral persistence is documented for immunosuppressed patients and may reach a whole year (Monrad et al. 2021; Sepulcri et al.

2021; Cunha et al. 2021). The reasons for immune system suppression in described cases are different. Mostly it is caused by B-cell depleted treatment, needed to cure hematological malignancies (Avanzato et al. 2020; Hueso et al. 2020; Reuken et al. 2021; Betrains et al. 2021; Monrad et al. 2021; Leung et al. 2022; Borges et al. 2021) or autoimmune disorders (Choi et al. 2020; Islam et al. 2021; Zabalza et al. 2021). The lack of humoral adaptive immunity certainly is positively associated with the longevity of COVID-19 (Fu et al., 2020).

Viral persistence is accompanied by intrahost evolution and accumulation of changes in the viral genome. Unfortunately, not all studies of long-term COVID-19 focus on dynamics of viral evolution and publish viral sequences (**Table 3.1**). Even fewer reports include the analysis of mutation effects on the fitness of the virus and the host immune system. Nevertheless, based on the available data, some conclusions can be drawn. The first one is that the number of accumulated changes in the viral genome dramatically varies between patients and is poorly associated with the duration of viral persistence. For example, in the case described by (Mukhina et al., 2022), the virus gained 5 SNPs over a period of 216 days, while in the case from (Choi et al., 2020) there were 29 SNPs and 2 deletions detected as a result of viral persistence during 154 days. Obviously, comparison of different cases may be biased by differences in technical details of the analysis. These may include differences in determination of duration of COVID-19, differences in detection of changes in viral genome from sequencing data, as well as differences in determination of the genome sequence of the ancestral virus, which became a source of the infection. Nonetheless, dramatic variability of evolutionary rate between different cases implies that individual features of a patient's immune system and used therapy largely determine the dynamics of viral evolution. Also individual mutations gained by the virus may affect mutation rate of the viral population in the particular host.

In several reports some additional analysis of viral evolution was done. In the study of Williamson et al., 2021 authors observed an increase of viral evolutionary rate after the injection of convalescent plasma. They explain this effect by viral escape from neutralizing antibodies, since mutations were shown to have such an effect in other studies. In another study of (C. Y. Lee et al. 2021) authors analyzed a huge cohort of 382 patients with hematological malignancies and prolonged COVID-19. They revealed that in a group with higher mortality and higher severity of COVID-10 there was an elevated dNdS.

Long-term COVID-19 is also repeatedly described for HIV positive individuals (Cele et al. 2021; Karim et al. 2021; Cunha et al. 2021). In general, the prevalence of SARS-CoV-2 cases among HIV infected individuals is the same as in the general population (Mirzaei et al. 2021; Guo et al. 2020). However, such an estimate does not take into account asymptomatic cases of COVID-19 and HIV infected individuals with unknown HIV status (Mirzaei et al. 2021). The nature of immunosuppression in HIV infected individuals is different from immunosuppression resulting from B-cell depleted therapy, but has some common features (Ambrosioni et al. 2021). HIV infection comes with the reduction of CD4 T cells, which are involved in activation of B cell response and production of neutralizing antibodies (Spinelli et al. 2021). Thus, the branch of humoral adaptive immunity is weakened in HIV infected individuals, as it happens after B-cell depleted therapy. It manifests itself in weaker and delayed response for various vaccines, such as pneumococcal, influenza and hepatitis B vaccines, in HIV infected individuals even on the background of ART therapy (Geretti and Doyle 2010). Much weaker response to SARS-CoV-2 vaccines, developed by Moderna and Pfizer, than in the general population was detected as well (Wang 2021). Severity of COVID-19 in HIV positive patients strongly depends on the level in CD4 T cell counts reduction and is less likely in patients receiving ART (Geretti and Doyle 2010; Tesoriero et al. 2021).

Thus, longevity of viral persistence in HIV infected individuals is comparable with B-cell depleted patients and can also reach almost a year (Karim et al. 2021). Two studies revealed that the virus acquired mutations, helping escape neutralizing antibodies (Cele et al. 2021; Karim et al. 2021). Moreover, part of acquired mutations in the patient, described in (Cele et al. 2021), matched mutations of VOCs, especially in the RBD region of S protein, which are known as variants increasing viral transmissivity. Thus, development of SARS-CoV-2 strains could be associated with the involvement of an HIV positive immune system. The fact that 25 out of the overall 38 millions HIV positive people live in sub-Saharan Africa, with the most of those not receiving ART, together with the fact that two of major SARS-CoV-2 strains appeared in South Africa, support this hypothesis.

**Table 3.1**. The summary of case reports of long-term COVID-19 in patients with various health conditions.

| Study | Cohort (# of patients) | Days | # of mutations | Comments | Immune escape |
|---|---|---|---|---|---|
| (Agarwal et al. 2020) | usual (851) | >= 28 | no data | 99 of 851 patients shed viral RNA after 4 weeks from initial diagnosis | no data |
| (Tillett et al. 2021) | usual (1) | 49 | 10 and 13 SNPs | reinfection | no data |
| (Gao and Wang 2015) | usual (22) | 50-112 | 0-3 SNPs | long-term COVId is associated with low viral load and decreased pathogenicity | no data |
| (Prado-Vivar et al. 2021) | usual (1) | 84 | 1 SNP in each infections | reinfection | no data |
| (Voloch et al. 2021) | usual (33) | ~18 on average | 0-7 SNPs | revealed common mutational profiles, associated with RdRp mutational error spectrums | no data |
| (Avanzato et al. 2020) | Chronic Lymphocytic Leukemia (1) | 70 | 12 SNPs; 3 del | shedding viral RNA uo to 105 days; CCP | no data |
| (Hueso et al. 2020; Betrains et al. 2021) | Diffuse Large B cell Lymphoma (5), Chronic Lymphocytic Leukemia (3), Mantle Cell Lymphoma (3), Follicular Lymphoma (3), Waldenström Macroglobulinemia (1), Marginal Zone Lymphoma (1), Multiple Sclerosis (1) | 7 - 83 | no data | anti-CD20 treatment (rituximab); intact T-cell immunity and lack of neutralizing antibodies; CCP | no data |

| | | | | | |
|---|---|---|---|---|---|
| (Reuken et al. 2021) | Follicular Lymphoma (1) | ~120 | 12 SNPs | anti-CD20 treatment (rituximab); CCP; elevated fraction of CD8 T cells and diminished fraction of CD4 T cells | no data |
| (Sepulcri et al. 2021) | Non-Hodgkin Lymphoma (1) | 238 | 18 amino-acid changing SNPs | anti-CD20 treatment (rituximab); CCP | no data |
| (Williamson et al. 2021) | Chronic Lymphocytic Leukemia (1) | 197 | 14 SNPs;<br>3 del | anti-CD20 treatment (rituximab); CCP; virus evolved rapidly between day 58 and 155 | no data, but mutations with known effect are detected (increasing transmissivity and escaping neutralizing antibodies) |
| (Monrad et al. 2021) | Chronic Lymphocytic Leukemia (1) | 333 | 8 SNPs and 3 del in Spike | CCP | escape from neutralizing antibodies |
| (Leung et al. 2022) | B-cell acute lymphoblastic leukemia (1) | 97 | 6 SNPs,<br>1 del | no CCP | no data |
| (C. Y. Lee et al. 2021) | 368 lymphoid malignancies | > 30 days | 18 patients sequenced | elevated dNdS in the group with higher mortality | no data |
| (Borges et al. 2021) | 1non-Hodgkin lymphoma | 6 months | 15 snp + 3 del | no CCP | no data |
| (Mukhina et al. 2022) | Mantle Cell Lymphoma (1) | 216 | 5 SNPs | anti-CD20 treatment (rituximab); no CCP | no data |
| (Choi et al. 2020) | Severe Antiphospholipid Syndrome (1) | 154 | 29 SNPs;<br>2 del | anti-CD20 treatment (rituximab); no CCP | no data |
| (Islam et al. 2021) | Metabolic Syndrome (1) | 72 | no data | no CCP | no data |
| (Kemp et al. 2020) | B-cell depleted (1) | 101 | > 30 SNPs | 23 sequenced time points over 101 days, little evolutionary change in first 65 | escape from neutralizing antibodies |

| | | | | | |
|---|---|---|---|---|---|
| | | | | days, but more after the start of CCP treatment | |
| (Khatamzas, Rehn, et al. 2021) | B-cell depleted (1) | 154 | 16 SNPs; 1 del | anti-CD20 antibody obinituzumab; CCP | no data |
| (Mulder et al. 2021) | Waldenström Macroglobulinemia (1) | 59 | 10 SNPs | anti-CD20 therapy, CCP, reinfection | no data |
| (Zabalza et al. 2021) | Multiple Sclerosis (48) | no data | no data | anti-CD20 treatment; patients after the treatment are less likely to generate antibody response | no data |
| (Weigang et al. 2021) | Kidney transplant recipient (1) | 105 | 34 SNPs, 2 del | immunosuppressive treatment (tacrolimus, mycophenolate mofetil, prednisone), CCP | escape from neutralizing antibodies |
| (Karim et al. 2021) | HIV (1) | 216 | 18 amino-acid changing SNPs, 2 del | antiretroviral treatment failure | escape from neutralizing antibodies |
| (Cunha et al. 2021) | usual (2), HIV (1) | 232 | ~20 SNPs | no data | no data |
| (Cele et al. 2021) | HIV (1) | 123 | 15 amino-acid changing SNPs, 1 del | ART therapy and switch to dolutegravir based therapy; many mutations match with mutations of VOCs, especially in RBD region of S protein | escape from neutralizing antibodies, including vaccine induced |

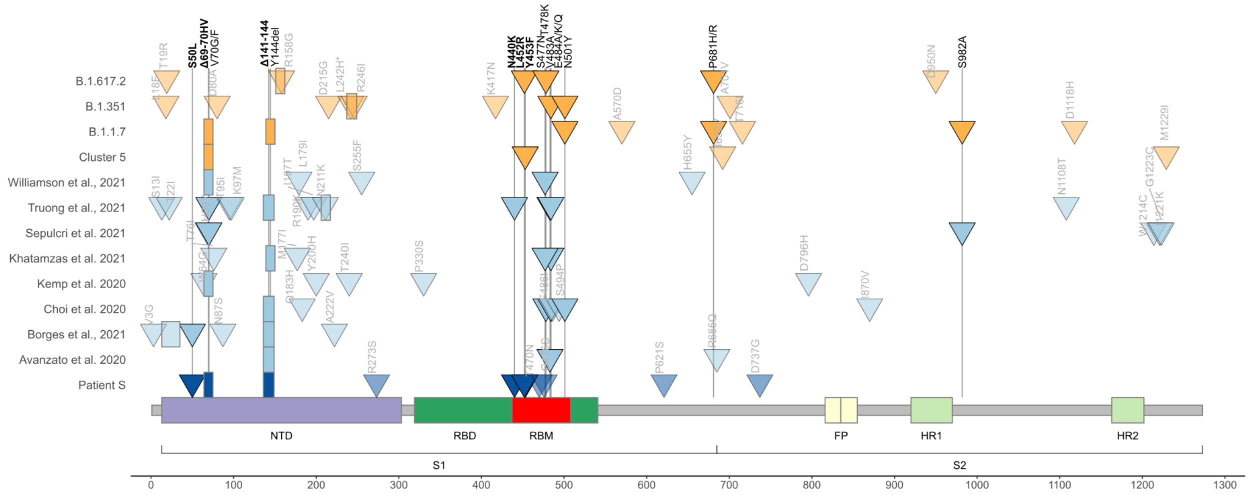


**Figure 2.1. Concordant origin of spike mutations in notable COVID-19 variants and reported cases of persistent COVID-19.** Shown are the locations of mutations in the amino acid sequence encoded by the spike gene. Rows, from top to bottom: VOCs Delta (B.1.617.2), Beta (B.1.351), Alpha (B.1.1.7); Cluster 5 variant; immunosuppressed individual with persistent infection for 290 days (Williamson et al., 2021); three patients with acute lymphoblastic leukemia who were persistently positive for SARS-CoV-2 (Truong et al., 2021); immunosuppressed individual treated with immunoglobulin (Sepulcri et al. 2021); immunosuppressed individual treated with convalescent plasma (Khatamzas et al. 2021); immunosuppressed individual treated with convalescent plasma (Kemp et al. 2020); immunosuppressed individual treated with Regeneron monoclonal antibody cocktail (Choi et al. 2020; only those mutations present at the final timepoint (T3, day 152) are shown); immunocompromised patient without convalescent plasma treatment (Borges et al., 2021); immunocompromised individual treated with convalescent plasma (Avanzato et al. 2020); immunosuppressed individual not treated with convalescent plasma or antibodies (patient S, this study). Triangles, point mutations; rectangles, deletions. Bright colors represent mutations observed in at least two studies. Mutations labeled on top in black were observed in multiple lineages/experiments, among those, mutations that are present in Patient S are highlighted with bold font.

# CHAPTER 3: MEMORY PERSISTENCE AND DIFFERENTIATION INTO ANTIBODY-SECRETING CELLS ACCOMPANIED BY POSITIVE SELECTION IN LONGITUDINAL BCR REPERTOIRES

In this Chapter we study the complex system of host-pathogen interaction from the side of the host. Being much simpler in their structure, pathogens have dramatically shorter generation times than their hosts, which enables them to develop new adaptive features relatively fast. Adaptive immunity possesses two mechanisms to respond promptly for such fast changes. The first mechanism encapsulates the whole idea of adaptive immunity: large pre-generated diversity of B and T cell receptors allows the inclusion of a new B or T cell clone in ongoing immune response at any moment. Thus any changes in a pathogen, hiding it from existing B and T responding clones will produce new antigens and evoke new clonal expansion.

The second mechanism is specific for B cell response. In addition to pre generated diversity of B cell receptors B cell clones can evolve during the interaction with the antigen in the process of affinity maturation. The structure of immune repertoires and phylogenies of B cell clonal lineages as its part contain the history of encounters with antigens. In this Chapter we focus on most abundant B cell clonal lineages, assigned from immunoglobulin heavy chain repertoires from memory B cells, plasmablasts, and plasma cells from peripheral blood collected from generally healthy volunteers at three time-points over the course of a year. By evolutionary and phylogenetic analysis of such lineages we study mechanisms, staying behind the intrahost adaptation of B cell immune response.

## Methods

### Donors, cells, and timepoints

Blood samples from six (4 males and 2 females) young and middle-aged donors (27, 27, 33, 33, and 39 y.o.) without severe inflammatory diseases, chronic or recent acute infectious diseases, or vaccinations were collected at three time-points (T1 - 0, T2 - 1 month, T3 - 12 months); donor details and the number cells collected for each time point and cell subset are provided in **Supplementary Table A-1**. Four donors suffered allergic rhinitis to pollen, and two also suffered from food allergy. Informed consent was obtained from each donor. The study was approved by the Ethical Committee of Pirogov Russian National Research Medical University, Moscow, Russia. At each time point, 18–22 mL of peripheral blood was collected in BD Vacuette tubes with EDTA. Peripheral blood mononuclear cells were isolated using Ficoll gradient density centrifugation. To isolate subpopulations of interest, cells were stained with anti-CD19-APC, anti-CD20-VioBlue, anti-CD27-VioBright FITC, and anti-CD138-PE-Vio770 (all Miltenyi Biotec) in the presence of FcR Blocking Reagent (Miltenyi Biotec) according to the manufacturer’s protocol, and then sorted using fluorescence-activated cell sorting (FACS; BD FacsAria III, BD Biosciences) into the following populations: memory B cells (Bmem; CD19$^{+}$ CD20$^{+}$ CD27$^{+}$ CD138$^{-}$, plasmablasts (PBL; CD20$^{–}$ CD19$^{Low/+}$ CD27$^{++}$ CD138$^{–}$), plasma cells (PL; CD20$^{–}$ CD19 $^{Low/+}$ CD27$^{++}$ CD138$^{+}$). For each donor at T1, one replicate sample of each cell subpopulation was collected. At T2 and T3, two replicate samples were collected ($50 \times 10^{3}$ to $100 \times 10^{3}$ Bmem, $1 \times 10^{3}$ to $2 \times 10^{3}$ PBL, $0.5 \times 10^{3}$ to $1 \times 10^{3}$ PL per sample).

### IGH cDNA libraries and sequencing

IGH cDNA libraries were prepared as described previously (Turchaninova et al. 2016) with several modifications. Briefly, we used a rapid amplification of cDNA ends (RACE) approach with a template-switch effect to introduce 5’ adaptors during cDNA synthesis. These adaptors contained both unique molecular identifiers (UMIs), allowing error-correction, and sample barcodes, allowing us to rule out potential cross-sample contaminations. In addition to a universal sequence for annealing the forward PCR primer, we also introduced a 5' adaptor during the reverse transcription (RT) reaction, which allowed us to avoid using multiplexed forward primers specific for V segments, thereby reducing PCR amplification biases. Multiplexed C-segment-specific primers were used for RT and PCR, allowing us to preserve isotype information. Prepared libraries were then sequenced with an Illumina HiSeq 2000/2500, (paired-end, 2 x 310 bp).

### Sequencing data pre-processing and repertoire reconstruction

Sample demultiplexing by sample-barcodes introduced in the 5' adapter and UMI-based error-correction were performed using MIGEC v1.2.7 software (Shugay et al. 2014). For further analysis, we used sequences covered by at least two sequencing reads. Alignment of sequences, V-, D-, J-, and C-segment annotation, and reconstruction of clonal repertoires were accomplished using MiXCR (Bolotin et al. 2015) with prior removal of the primer-originated component of the C-segment. We defined clonotypes as a unique IGH nucleotide sequence starting from the framework 1 region of the V segment to the end of the J segment, and taking into account isotype. Using TIgGER (Gadala-Maria et al. 2015) software, we derived an individual database of V gene alleles for each donor and realigned all sequences for precise detection of hypermutations. For analysis of general repertoire characteristics (isotype frequencies, SHM levels, CDR3 length, IGHV gene usage, and repertoire similarity metrics) we used samples covered by at least 0.1 cDNA molecules per cell for Bmem, and at least 5 cDNA per cell for PBL and PL.

### Assignment of clonal lineages

Change-O v0.4.4 (Gupta et al. 2015) was utilized to assign clonal groups, defined as groups of clonotypes with the same V segment, CDR3 length, and at least 85% similarity in CDR3 nucleotide sequence. Before clonal group assignment, we excluded all clonotypes with counts equal to 1. Clonal groups represent observed subsets of clonal lineages originating from a single BCR ancestor, so for simplicity, we use the term ‘clonal lineages’. To study evolutionary dynamics of clonal lineages, we joined all replicas, three time-points (T1, T2, and T3), and cell subsets for each patient into a single dataset and excluded clonotypes that were presented by a single UMI. Phylogenetic analysis was performed on four patients for whom we had samples at all time-points, and on clonal lineages containing at least 20 unique clonotypes as in (Nourmohammad et al. 2019).

### Clusterization of clonal lineages in HBmem and LBmem clusters

We performed principal component analysis on six scaled variables of clonal lineage composition: fractions of Bmem, PBL, and PL, and fractions of IgM, IgG, and IgA. The IgE isotype was not detected in clonal lineages involved in phylogenetic analysis, so we did not include it as a variable. HBmem and LBmem clusters were defined using the K-means clustering algorithm.

### Metric of persistence of clonal lineages

We estimated the frequency of a clonal lineage in the repertoire at a given time-point as the ratio of the number of unique clonotypes in the clonal lineage detected at this time-point to the overall number of unique clonotypes detected at this time-point. If the clonal lineage was not detected at some time-point, we assigned its frequency to pseudocount, as it would be a single clonotype detected from this time-point. To estimate persistence of clonal lineage frequency in the repertoire over time we defined the persistence metric:

$$P = \frac{1}{\frac{1}{2}\left(\frac{f_{max}}{f_i} + \frac{f_{max}}{f_j}\right)},$$

where $f_{max}$ is the maximum frequency of the clonal lineage in the three time-points and $f_{i,j}$ are its frequencies in the other two (**Figure 3.2D**). Persistence is equal to 1 if the frequency remains consistent at all three time points. If a clonal lineage was detected just once in the experiment and frequencies at other two time points were assigned to pseudocounts, the persistence approaches zero.

### Reconstruction of clonal lineage germline sequence

We used MiXCR-derived reference V, D, and J segment sequences to reconstruct IGH germline sequences for each clonal lineage, concatenating only those sequence fragments which were present at CDR3 junctions of original MiXCR-defined clonotypes. Thus, random nucleotide insertions were disregarded, making them appear as gaps in the alignment of lineage clonotypes with the germline sequence. We excluded them from all parts of the phylogenetic analysis where germline sequence was required.

### Reconstruction of clonal lineage phylogeny and MRCA

For phylogenetic analysis of clonal lineages, we aligned clonotypes with reconstructed germline sequences using MUSCLE version 3.8.31 with 400 gap open penalty (Edgar 2004). Next, we reconstructed the clonal lineage's phylogeny with RAxML version 8.2.11, using the GTRGAMMA evolutionary model and germline sequence as an outgroup, and computed marginal ancestral states (Stamatakis 2014). The ancestral sequence of the node closest to the root of the tree, represented by the germline sequence, is the MRCA of the sampled clonotypes. It can match the germline sequence or differ by some amount due to SHM, reflecting the starting point of subsequent evolution of observed clonotypes. This allowed us to distinguish between SHMs fixed in the clonal lineage on the way from the germline sequence to the MRCA (G-MRCA SHMs) versus polymorphisms within the observed part of lineage. The G-MRCA p-distance in **Figure 3.3B** was measured as a fraction of diverged positions between germline and MRCA sequences.

**McDonald-Kreitman (MK) test**

The (MK) test is designed to detect the effects of positive or negative selection on population divergence from another species or its ancestral state (McDonald and Kreitman 1991). It is based on the comparison of ratios of nonsynonymous to synonymous substitutions observed in diverged and polymorphic sites, and estimates the fraction of diverged amino acid substitutions fixed by positive selection:

$$\alpha = 1 - \frac{P_n}{P_s} \cdot \frac{D_s}{D_n},$$

where $P_n$ and $P_s$ respectively represent nonsynonymous and synonymous polymorphisms, and $D_n$ and $D_s$ respectively represent nonsynonymous and synonymous divergences fixed in the population. Under neutral evolution, nonsynonymous and synonymous changes are equally likely to be fixed or appear in the population as polymorphisms, so $\frac{D_n}{D_s} = \frac{P_n}{P_s}$ and $\alpha = 0$. Positive selection favors adaptive nonsynonymous changes to be fixed, and increases $\frac{D_n}{D_s}$ relative to $\frac{P_n}{P_s}$, resulting in $\alpha > 0$. Negative selection has the opposite effect and produces $\alpha < 0$.

To detect selection in the origin of clonal lineages, we considered G-MRCA SHM as divergent changes, and the remaining SHM in a clonal lineage after the MRCA as polymorphic ones (**Figure 3.4A**). If we observed different nucleotides in the germline sequence and MRCA at a site that was also polymorphic, we considered it as divergent only if the germline variant was not among the polymorphisms (**Supplementary Table A-2**, examples of codons q and r). Codons with unknown germline state were excluded from the MK test (**Supplementary Table A-2**, example of codon j). To perform the MK test on joined HBmem or LBmem cluster variation, we summed variation of all clonal lineages of the same cluster in each category ($D_n$,$D_s$, $P_n$, $P_s$). Calculations of α of distinct clonal lineages for comparison of its distributions between two clusters were complicated by zero G-MRCA distance in some clonal lineages, mostly belonging to the HBmem cluster. We dealt with this using three approaches, presented in **Supplementary Table A-3**. In the first, we added pseudocounts to $D_n$ and$D_s$ in each clonal lineage, so that for clonal lineages with zero G-MRCA distance, $\frac{D_n}{D_s} = 1$. In the second, we excluded clonal lineages with zero G-MRCA distance from the analysis, still adding pseudocounts to $D_n$ and $D_s$ in each clonal lineage in cases where the G-MRCA distance consists of just one nonsynonymous or synonymous substitution. In the third, we compared only those clonal lineages that had at least one nonsynonymous and at least one synonymous substitution on the G-MRCA branch. We also calculated the MK test on joined variation for all types of exclusion criteria to check its robustness; however, there is no need to exclude clonal lineages in the case of the joined test (**Supplementary Table A-3**). In the first approach clonal lineages with zero G-MRCA distance always produced negative α and biased median α to negative values as well. Medians of α in the second and third approaches were more consistent with results of the test on joined variation. However, in the third approach, the filter excluded most of the HBmem cluster, and so in the main test we presented results of the second approach (**Figure 3.4B**). To check the significance of deviation of α from neutral expectations, we used an exact Fisher test as in the original MK pipeline (McDonald and Kreitman 1991).

**πNπS**

To calculate πNπS we identified SHMs in each clonal lineage relative to the reconstructed MRCA sequence. In multiallelic sites (with multiple SHMs observed, see codon *i* in **Supplementary Table A-2** as an example) we considered each variant as an independent SHM event. πN and πS were calculated as the number of nonsynonymous and synonymous SHMs in a clonal lineage, normalized to the number of nonsynonymous and synonymous sites in the MRCA sequence respectively. The resulting πNπS value is the ratio between πN and πS:

$$\pi N \pi S = \frac{N}{N_S} : \frac{S}{S_S}$$

where $N$ and $S$ are the number of nonsynonymous or synonymous SHMs, respectively, observed in the clonal lineage and $N_S$ and $S_S$are the number of nonsynonymous or synonymous sites, respectively, in the MRCA sequence of the clonal lineage, calculated as in (Gojobori 1986).

**Site frequency spectrum**

Site frequency spectrum (SFS) reflects the distribution of SHM frequencies in the clonal lineage. We calculated the frequency of each SHM as a number of unique clonotypes carrying the SHM relative to the overall number of unique clonotypes in the lineage. To visualize SFS, we binned SHM frequencies into 20 equal intervals from 0 to 1 with a step size of 0.05, and counted SHM density in each bin as the number of SHMs in a given frequency bin normalized to the overall number of SHMs detected in the lineage. To obtain the cluster average SFS, we took the mean of clonal lineages of the same cluster in each frequency bin.

**Normalized πNπS in bins of SHM frequencies**

To compare ratios of nonsynonymous and synonymous SHMs of different frequencies between two clusters, we calculated normalized πNπS in bins of SHM frequency. For this purpose we used a smaller number of frequency bins (0; 0.2; 0.4; 0.6; 0.8; 1) to reduce the probability of bins without observed SHMs. To deal with the remaining empty bins, we added pseudocounts to nonsynonymous and synonymous SHMs in each frequency bin. Thus, normalized πNπS in the $i$-th SHM frequency bin was calculated as follows:

$$\text{normalized } \pi N \pi S = \frac{(N_i+1)/(\sum_{i=1}^{5} N_i+5)/N_S}{(S_i+1)/(\sum_{i=1}^{5} S_i+5)/S_S}$$

where $N_i$ and $S_i$ are the number of nonsynonymous and synonymous SHMs, respectively, in the $i$-th frequency bin, $\sum_{i=1}^{5} N_i$ and $\sum_{i=1}^{5} S_i$ are respectively the overall number of nonsynonymous and synonymous SHMs observed in the clonal lineage (the sum of SHMs in all frequency bins), and

$N_S$ and $S_S$ are the number of nonsynonymous and synonymous sites respectively in the MRCA sequence of the clonal lineage, calculated as in (Gojobori 1986). To compare distributions of normalized πNπS between two clusters of clonal lineages in the five frequency bins, we used the Mann-Whitney test with Bonferroni-Holm multiple testing correction.

### Data analysis and visualization

All analysis was performed using R language (R Core Team 2018) and visualized with the ggplot2 package (Ginestet 2011). Ggtree package was used to visualize phylogenetic trees of clonal lineages (Yu et al. 2017). The code is available at https://github.com/EvgeniiaAlekseeva/Clonal_group_analysis.

## Results

### Definition of the most abundant B cell clonal lineages for phylogenetic analysis

We collected peripheral blood from six healthy donors at three time-points, where the second sample was collected one month after the first, and the third was collected 11 months after that (**Figure 3.1, Supplementary Table A-1**). These samples were subjected to fluorescence-activated cell sorting to isolate memory B cells (Bmem: CD19 $^{+}$ CD20 $^{+}$ CD27 $^{+}$ ), plasmablasts (PBL: CD19 $^{low/+}$ CD20 $^{-}$ CD27 $^{high}$ CD138 $^{-}$ ) and plasma cells (PL: CD19 $^{low/+}$ CD20 $^{-}$ CD27 $^{high}$ CD138 $^{+}$ ). Most of the cell samples were collected and processed in two independent replicates. For each cell sample, we obtained full-length IGH clonal repertoires were obtained using a 5’-RACE-based protocol, which allows unbiased amplification of full-length IGH variable domain cDNA while preserving isotype information, with subsequent unique molecular identifier (UMI)-based sequencing data normalization and error correction (Turchaninova et al. 2016; Shugay et al. 2014).

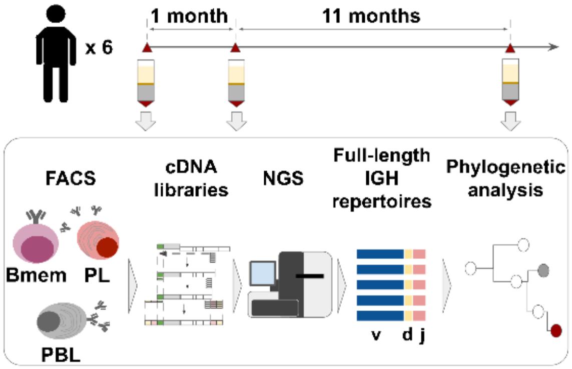


**Figure 3.1 Study design.** Peripheral blood from six donors was sampled at three time points: T1 - initial time point, T2 - 1 month and T3 - 12 months after the start of the study. The figure is adapted from (Mikelov et al., 2022).

Extensive comparative analysis of immune repertoires of B cell subsets on the general level can be found in our publication (Mikelov et al. 2022). The part of this work, constituting the third chapter of this thesis, is devoted to the phylogenetic analysis of the most abundant B cell clonal lineages. Since we aimed to study associations between lineage dynamics and its evolutionary regime, for this part of analysis immune repertoires of four of six donors from whom samples were collected at each of the three time-points we used. We defined the most abundant clonal lineages as lineages consisting of at least 20 unique clonotypes from the corresponding donor. On average, these clonal lineages covered 3.4% of a given donor’s repertoire (**Supplementary Figure A-1A**), and we identified 190 such lineages across the four donors.

### Temporal dynamics of clonal lineages are associated with cell subset composition

First we asked how B cell subsets and isotypes were presented among these most abundant clonal lineages. Clonal lineages were mostly composed of memory B-cell clonotypes of non-switched isotype IgM either were largely composed of ASCs, and enriched in IgG and IgA clonotypes (**Supplementary Figure A-1B**). To investigate the nature of such bimodal distribution and perform comparative analysis of these two types of clonal lineages we divided them into two large clusters using k-means clustering algorithm, based on the proportion of cell subsets and BCR isotypes represented (**Figure 3.2A:B, Supplementary Figure A-2A**). The more abundant HBmem cluster included 138 clonal lineages, and was mostly composed of memory B-cell clonotypes of non-switched isotype IgM. Conversely, the smaller LBmem cluster

(52 clonal lineages) was more diverse and largely composed of ASCs, and enriched in IgG and IgA clonotypes. The average size of clonal lineages (*i.e.*, the number of unique clonotypes per lineage) did not differ between the HBmem and LBmem clusters (**Supplementary Figure A-2B**), and both clusters were presented in repertoires of all donors (**Supplementary Figure A-2C**).

Next we tracked the abundance of each clonal lineage in the repertoire across each time-point. The two clusters of lineages demonstrated different temporal behavior; while HBmem groups were quite stable over time, LBmem lineages had a burst of increased frequency at one of the time points (**Figure 3.2C**). To compare the temporal stability of clonal lineages, we defined the lineage persistence metric, which equals 1 when a clonal lineage was equally frequent at all three time-points and is close to 0 when it was detected at just one time-point (**Figure 3.2D**). Persistence of a clonal lineage was strongly associated with its composition (**Figure 3.2E:F**). Clonal lineages enriched with clonotypes or with the IgM isotype — including all HBmem lineages — were more likely to persist through time. Conversely, lineages with larger proportions of ASCs or IgG/IgA isotypes, including most LBmem lineages, tended to have lower persistence, with a burst of increased frequency at one particular time-point. The time-point of increased LBmem frequency varies among donors (**Supplementary Figure A-2D**). The frequencies of clonal lineages were highly correlated among replicate samples, and the persistence of a clonal lineage was not associated with its size (**Supplementary Figure A-2E:F**), indicating that differences in persistence cannot be attributed to clonotype sampling noise.

Besides their higher persistence, the HBmem lineages were enriched in clonotypes detected at multiple time-points (**Supplementary Figure A-2G**), indicating that persistent clonal lineages are supported by persistent clonotypes. Furthermore, 29.7% of the HBmem cluster was represented by public clonal lineages shared between at least two donors, compared to 3.8% for the LBmem cluster. The only two shared LBmem lineages had atypically high persistence, which made them more similar to HBmem (**Figure 3.2G**).

Thus we observed two types of clonal lineages, representing different stages of immune response: persisting memory with unswitched IgM isotype (HBmem) and responding lineages with rapidly increased frequency, producing IgG or IgA antibodies (LBmem).

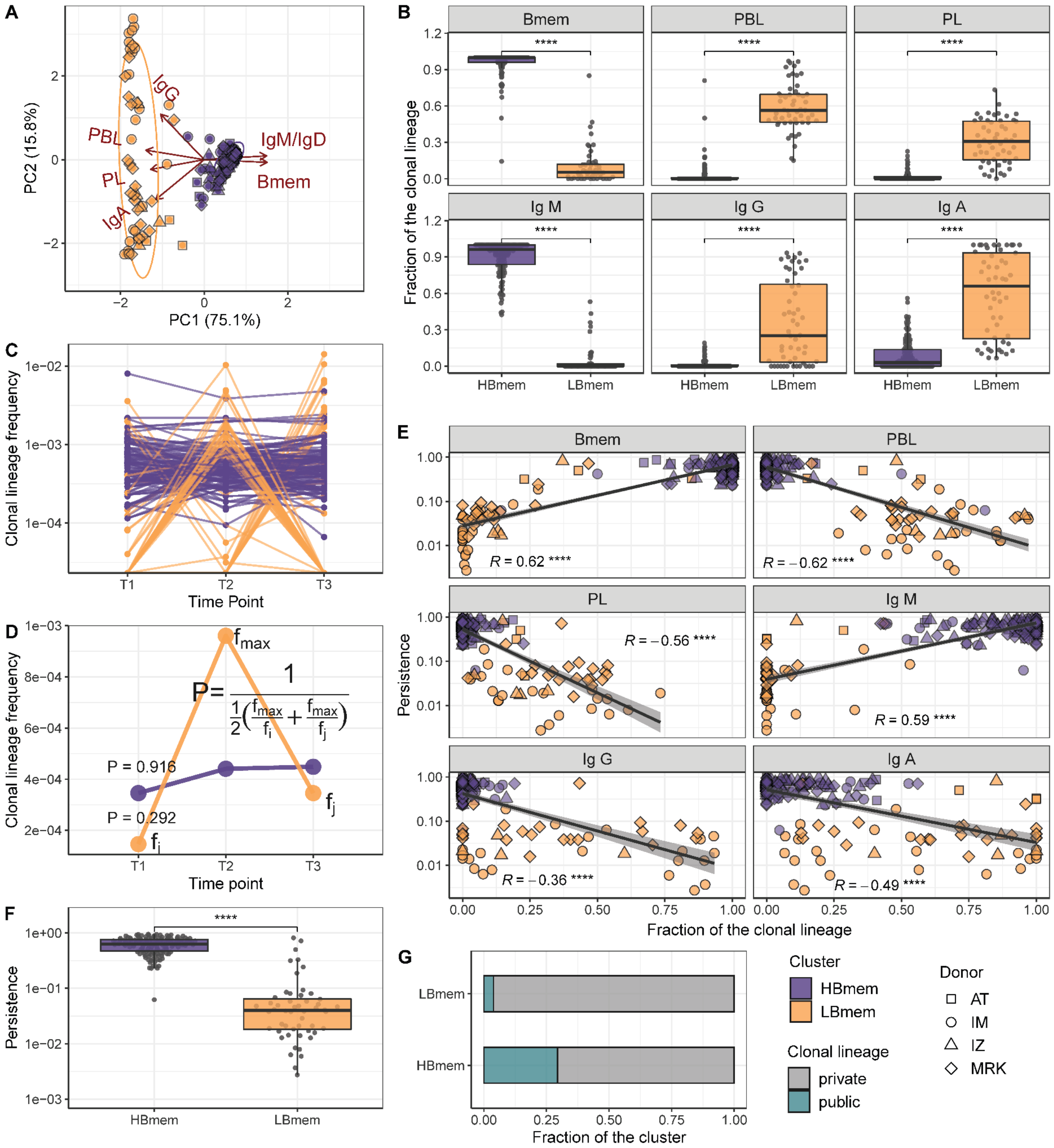


**Figure 3.2 Temporal dynamics and composition of clonal lineages. A**: Principal component analysis (PCA) of clonal lineage composition: proportions of Bmem, PBL, and PL as well as proportions of isotypes. Arrows represent projections of the corresponding variables onto the two-dimensional PCA plane, with lengths reflecting how well the variable explains the variance of the data. The two principal components (PC1 and PC2) cumulatively explain 90.9% of the variance. Clonal lineages assigned by the k-means algorithm to different clusters are shown in

different colors. **B**: Proportion of clonotypes from the various cell subset or isotypes in clonal lineages falling into the HBmem or LBmem clusters; **C**: Dynamics of clonal lineage frequency, defined as the number of clonotypes in a lineage divided by the total number of clonotypes detected at a given time-point. Each line connects points representing a unique clonal lineage (N=190). **D**: Schematic representation of how we calculated clonal lineage persistence. $f_{max}$ is the maximum clonal lineage frequency among the three time-points, and $f_{i,j}$ are the frequencies at the remaining two time-points. **E**: Spearman's correlation between persistence of a clonal lineage and proportions of its clonotypes associated with a given B cell subset or isotype. **F**: Comparison of persistence between HBmem and LBmem. **G**: Fraction of public clonal lineages significantly differs in HBmem (41 out of 97) and LBmem clusters (2 out of 50), Two-proportions Z test: $p \leq 10^{-5}$. Statistical significance for B and F is calculated by the two-sided Mann-Whitney test. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 10^{-3}$, **** = $p \leq 10^{-4}$.

### LBmem clonal lineages could arise from HBmem clonal lineages

The evolutionary past of a clonal lineage can be described by inferring the history of accumulation of SHMs leading to individual clonotypes—*i.e.*, by reconstructing the phylogenetic tree of the clonal lineage. The initial germline sequence of each clonal lineage partially matches the germline VDJ segments, and can be reconstructed in a manner corresponding to the root of the phylogenetic tree of this lineage (see Methods). However, the first node of the phylogenetic tree (green diamond in **Figure 3.3A**), the most recent common ancestor (MRCA) of the sampled part of the lineage, can be different from the inferred germline sequence. These differences, referred to as the G-MRCA distance, correspond to SHMs accumulated during the evolution of the clonal lineage prior to divergence of the observed clonotypes. The G-MRCA distance depends on how clonotypes of the tree were sampled. Sampling of clonotypes regardless of their position on the tree results in a low G-MRCA distance (**Figure 3.3A**, top panel), while sampling just those clonotypes belonging to a particular clade can conceal early stages of lineage evolution and thus result in a large G-MRCA distance (**Figure 3.3A**, bottom panel).

The G-MRCA distance was on average five-fold higher in LBmem clonal lineages compared to HBmem (median = 0.044 vs. 0.008, **Figure 3.3B**). This means that even though nearly all the evolution of an HBmem clonal lineage leaves a trace in the observed diversity of that lineage

(**Figure 3.3D, G**), the sequence variants of an LBmem lineage typically result from divergence of an already-hypermutated clonotype (**Figure 3.3E, H**). In most (38 out of 52) LBmem lineages, some Bmem clonotypes were observed at the time-point preceding expansion. Moreover, clonotypes of LBmem lineages are typically characterized by lower pairwise divergence compared to that in HBmem lineages (median = 0.11 vs 0.13, **Figure 3.3C**). Together with the burst-like dynamics characteristic of LBmem lineages (**Figure 3.2F**), this implies that LBmem lineages may represent recent, rapid clonal expansion of preexisting memory.

Based on these results and the compositional features of the two clusters, we further hypothesized that LBmem clonal lineages may arise from reactivation of pre-existing memory cells belonging to the HBmem cluster. In search of examples of such a transition, we examined all clonal lineages that were persistent but included ASC clonotypes. We found one clear example of a transition from HBmem to LBmem state in the evolutionary history of a clonal lineage (**Figure 3.3F, I**). While the MRCA of this lineage nearly matched the germline sequence, all ASC clonotypes were grouped in a single monophyletic clade (sublineage), such that its ancestral node was remote from the MRCA. The ASC sublineage demonstrated all features characteristic of LBmem, including predominance of IgG and IgA isotypes, low persistence, and low clonotype divergence. Conversely, the remainder of the clonal lineage had features of HBmem: predominance of IgM, high persistence, and high levels of clonotype divergence. Position of ASC sublineage on a distant node from the root of the tree indicates gradual accumulation of SHMs, differing ASC sublineage from the remaining clonotypes. This fact together with the similarity of CDR3 regions of lineage clonotypes (**Supplementary Figure A-3**) give a reason to conclude that the ASC sublineage has the same origin as the remaining part of the tree with features of HBmem cluster.

To summarize, we observed that LBmem lineages had low level of clonotype divergence and the large distance of lineage's ancestor from the germline sequence, assuming recent origin from a mature clonotype. The temporal dynamics of LBmem, detection of Bmem clonotypes at the time point prior to the LBmem lineage expansion, and the relationship between HBmem and LBmem on a clonal lineage level, together suggest that LBmem expansions may represent the result of reactivation of pre-existing memory.

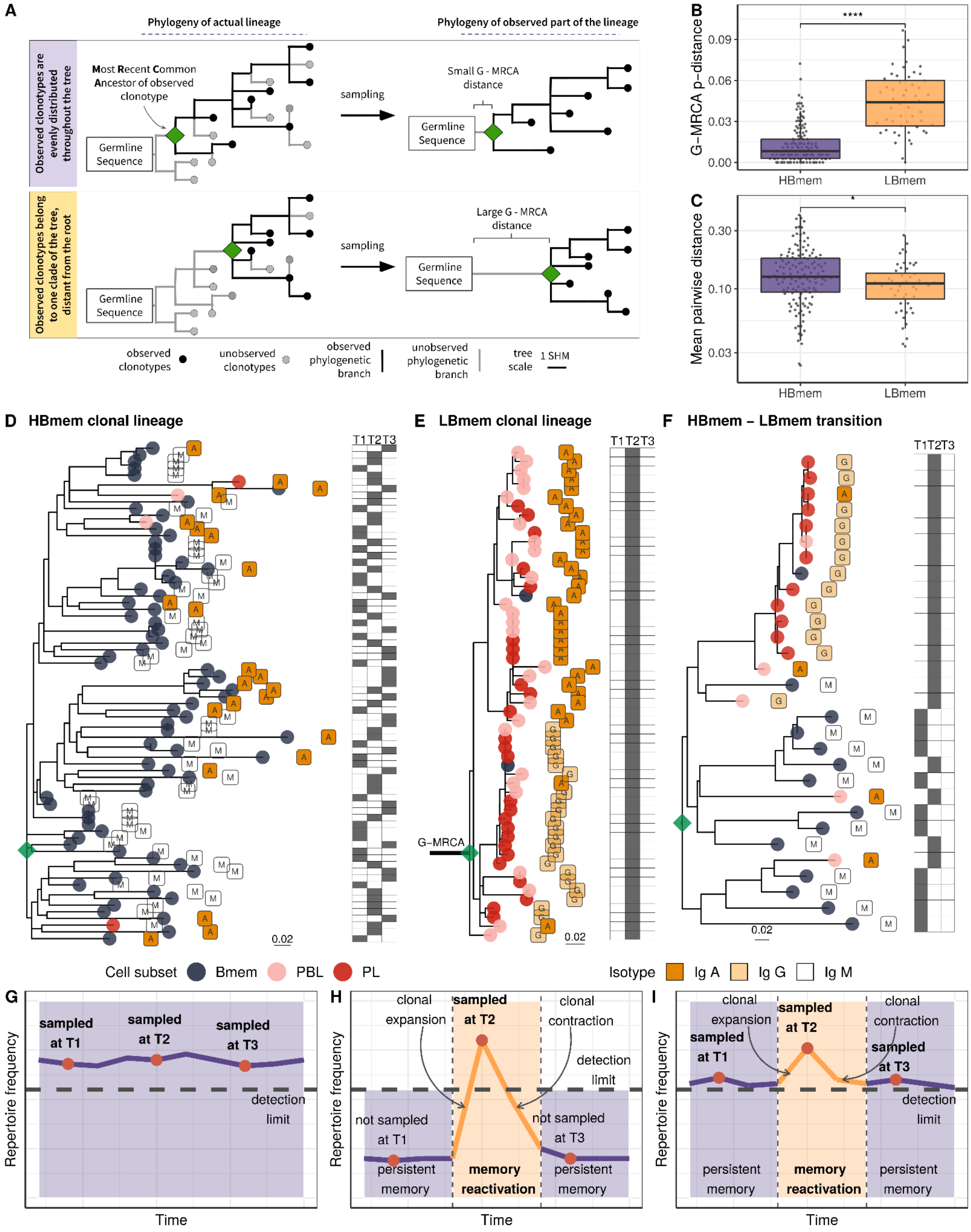


**Figure 3.3. Phylogenetic history of HBmem and LBmem clonal lineages. A**: A schematic illustration of how the distances between the germline sequence and the MRCA of a clonal

lineage (G-MRCA distance) vary depending on which subset of clonotypes is sampled: a sample uniform with regard to the position on the tree (top panel), or only those belonging to a particular clade of the tree (bottom panel). **B**: Comparison of G-MRCA p-distance (*i.e.*, the fraction of differing nucleotides) for HBmem and LBmem lineages. **C**: Mean pairwise phylogenetic distance (*i.e.*, the distance along the tree) between clonotypes of the same lineage for HBmem and LBmem clusters. **D–F**: Representative phylogenetic trees for clonal lineages belonging to HBmem (D), LBmem (E), and an example of HBmem-LBmem transition (F). The LBmem sublineage in F is nested deep in the phylogeny of the memory clonotypes, and is not characterized by a particularly long ancestral branch, indicating that it is not an artifact of clonal lineage assignment. Circles correspond to individual clonotypes, with the cellular subset indicated by color, and the isotype by label. Tables at right indicate the presence or absence of the corresponding clonotype at each time-point. The G-MRCA distance is indicated with a thick line. **G–I**: Schematic representation of the hypothetical dynamics of relative size for clonal lineages represented in D, E, and F, respectively. Significance for B and C was obtained by the two-sided Mann-Whitney test. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 10^{-3}$, **** = $p \leq 10^{-4}$.

**Reactivation of LBmem clonal lineages is driven by positive selection**

Having shown that the LBmem lineages likely originate from clonal expansion of pre-existing memory, we further compared the contribution of positive (favoring new beneficial SHMs) and negative (preserving the current variant) selection between the LBmem and HBmem clusters. Since we observed only one clear example of an HBmem-LBmem transition (**Figure 3.3F, Supplementary Figure A-3**), we could not claim with certainty that LBmem lineages always emerge from preexisting HBmem lineages rather than from some other memory type. Still, we were able to study LBmem reactivation by comparing differences in substitution patterns at the origin of HBmem and LBmem clusters. We reasoned that the G-MRCA distance of an HBmem lineage contains mutations fixed by primary affinity maturation after the initial lineage activation. In contrast, the G-MRCA distance of an LBmem lineage contains both mutations arising during primary affinity maturation and subsequent changes occurring later in the evolution of the lineage. Differences in the characteristics of the G-MRCA mutations between clusters are therefore informative of the process prior to the observed expansion of LBmem lineages.

To assess selection at the origin of the HBmem and LBmem lineages, we measured the divergence of nonsynonymous sites relative to synonymous sites (*i.e.*, the DnDs ratio). Classically, DnDs > 1 is interpreted as evidence for positive selection. However, DnDs > 1 is rare, because the signal of positive selection is usually swamped by that of negative selection. In the McDonald-Kreitman (MK) framework, positive selection is instead revealed by excessive nonsynonymous divergence relative to nonsynonymous polymorphism (*i.e.,* DnDs > PnPs; see Methods and **Supplementary Table A-2** for examples), under the logic that advantageous changes contribute more to divergence than to polymorphism (McDonald and Kreitman 1991). The fraction of adaptive nonsynonymous substitutions (α) can then be estimated from this excess. We designed an MK-like analysis, comparing the relative frequencies of nonsynonymous and synonymous SHMs at the G-MRCA branch (equivalent to divergence in the MK test) to those in subsequent evolution of clonal lineages (equivalent to polymorphism in the MK test; **Figure 3.4A**, see Methods).

In both the HBmem and LBmem clonal lineages, we observed a higher ratio of nonsynonymous to synonymous SHMs in the G-MRCA branches compared to subsequent tree branches, meaning that a fraction of SHMs acquired by MRCA was further fixed by positive selection. However, this fraction was higher in LBmem lineages (Fisher's exact test: $\alpha = 0.58$ and 0.65, $p < 10^{-6}$ and $< 10^{-15}$ in HBmem and LBmem, respectively). α of distinct clonal lineages was also generally higher in LBmem than in HBmem (median $\alpha = 0.57$ vs $\alpha = 0.18$, **Figure 3.4B**), showing that positive selection more frequently preceded expansion of LBmem than HBmem lineages. The observation of excess α in the LBmem cluster compared to HBmem was robust to the peculiarities of the MK analysis (**Supplementary Table A-3**). The higher α for LBmem compared to HBmem implies that a larger fraction of SHMs was positively selected in LBmem clonal lineages before their expansion. This excess of advantageous SHMs in ancestors of LBmem lineages together with previous observations that LBmem lineages likely originate from reactivated memory suggests that reactivation was coupled with new rounds of affinity maturation.

### Subsequent evolution of LBmem clonal lineages is affected by negative and positive selection

Next, we considered the effects of selection on HBmem and LBmem clusters following their divergence from their MRCAs — *i.e.*, in the subsequent evolution of a clonal lineage leading to the diversity of the observed clonotypes. We calculated the per-site ratio of nonsynonymous and synonymous SHMs (πNπS) among those that originated after the MRCA. The πNπS of both clusters was < 1 (**Figure 3.4C**). This deficit of nonsynonymous SHMs indicates negative selection in the observed part of clonal lineage evolution. The πNπS ratio was lower in the LBmem cluster, indicating stronger negative selection.

To examine the selection affecting these post-MRCA SHMs in more detail, we studied the frequency distribution of SHMs in individual lineages, or their site frequency spectra (SFS) (Nielsen 2005; Neher, Kessinger, and Shraiman 2013; Nei and Kumar 2000; Horns et al. 2019; Nourmohammad et al. 2019) (**Figure 3.4A**). SFS reflects the effect of selection on these SHMs. Deleterious SHMs are held back by negative selection, so that their frequency in the lineage remains low. By contrast, positive selection favors the spread of adaptive SHMs, increasing their frequency. Therefore, negative selection biases the SFS towards low frequencies, and positive selection, towards high frequencies. For each clonal lineage, we reconstructed the SFS of the SHMs accumulated since divergence from MRCA (**Figure 3.4A**), and then averaged these SFSs within the HBmem and LBmem clusters. A larger proportion of the LBmem SFS corresponds to high frequencies compared to HBmem (**Figure 3.4D**), indicating weaker negative and/or stronger positive selection in LBmem SFS.

To distinguish between these selection types, we calculated the proportion of the SFS distribution falling into each frequency bin for nonsynonymous SHMs, and divided it by the same value for synonymous SHMs (normalized πNπS; see Methods, **Figure 3.4E**). The inter-cluster differences in the normalized πNπS in low-frequency bins were generally reflective of negative selection, while the differences in the high-frequency bins were reflective of positive selection. Normalized πNπS was significantly higher in the high-frequency (>60%) bins of SHMs in LBmem clonal lineages. This indicates that for LBmem, those nonsynonymous changes that were not removed by negative selection reached high frequencies more often than in HBmem. In total, these data indicate that a fraction of nonsynonymous mutations accumulated by LBmem lineages were adaptive. We thus observed that reactivation of LBmem lineages is coupled with strengthening of both types of selection: positive on the G-MRCA branch, and both positive and negative during subsequent clonal lineage expansion. This pattern is most likely evidence of new rounds of affinity maturation, which result in the acquisition of new advantageous changes and preserve the resulting BCRs from deleterious ones. HBmem, in contrast, evolved more neutrally under weaker negative selection, suggesting absence of antigen challenge during the observation period (**Figure 3.4F**).

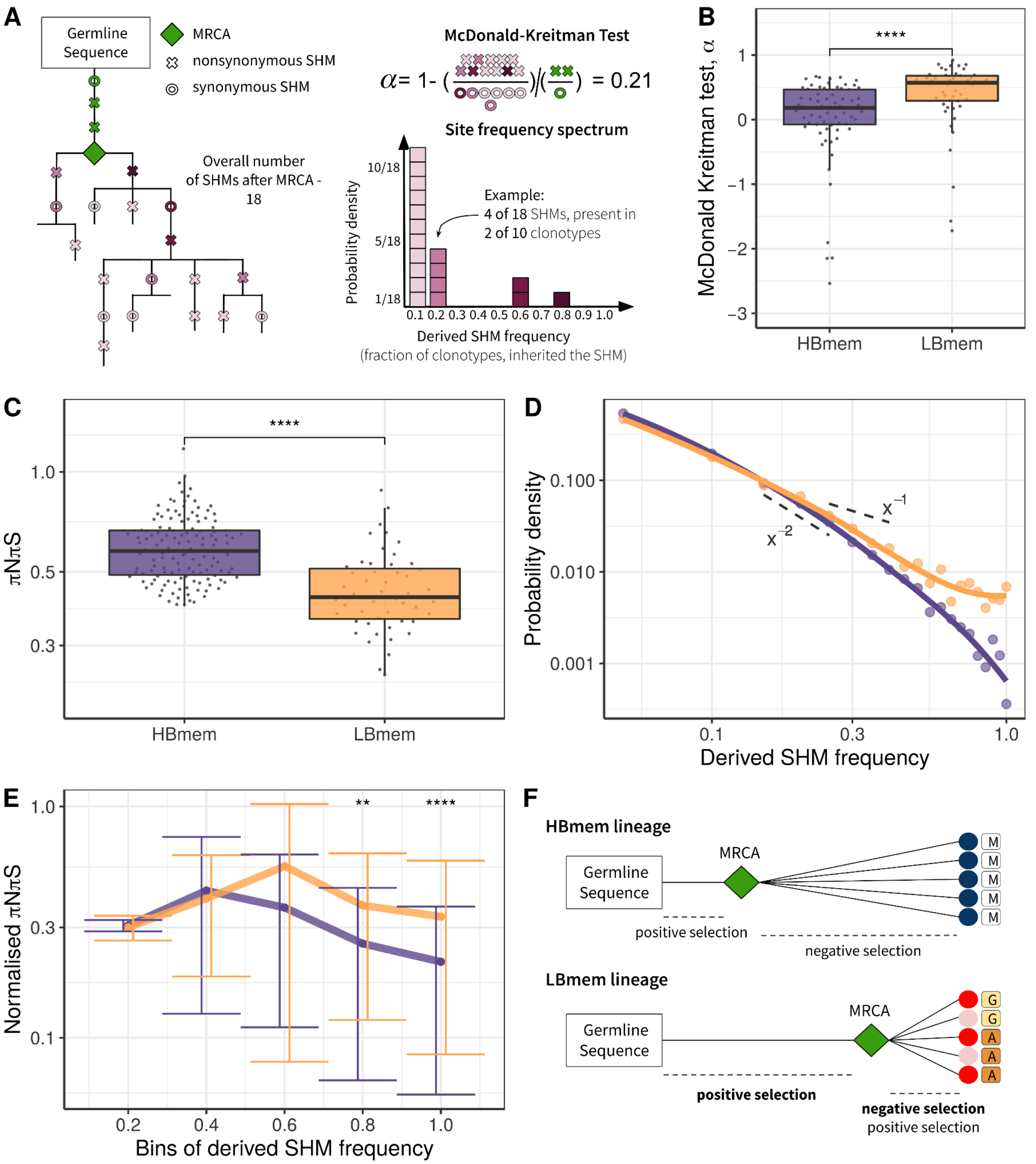


**Figure 3.4 Signatures of positive and negative selection in HBmem and LBmem clusters. A**: Schematic of the McDonald-Kreitman (MK) test and site frequency spectrum (SFS) concept. **B**: MK estimate of the fraction of adaptive non-synonymous changes (α) between germline and MRCA in HBmem and LBmem clonal lineages. Only lineages with nonzero G-MRCA distance are included. N = 68 for HBmem, 49 for LBmem, see Supplementary Table A-3); **C**:

Comparison of mean pairwise πNπS of HBmem and LBmem lineages. **D**: Averaged SFS for HBmem and LBmem clonal lineages. The two dashed lines correspond to $f(x)=x^{-1}$, which is the expected neutral SFS under Kingman's coalescent model (Kingman 1982), and $f(x)=x^{-2}$. **E**: Comparison of normalized πNπS for HBmem and LBmem clonal lineages in various SHM frequency bins. The number of polymorphisms in each bin is normalized to the overall number of polymorphisms in a corresponding clonal lineage. **F**: Scheme summarizing features of HBmem and LBmem clonal lineages. Comparisons in B, C, and E were performed by two-sided Mann-Whitney test, with Bonferroni-Holm multiple testing correction in E. * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 10^{-3}$, **** = $p \leq 10^{-4}$.

## Discussion

Using advanced library preparation technology, we performed a longitudinal study of BCR repertoires of the three main antigen-experienced B cell subsets — memory B cells, plasmablasts, and plasma cells — from peripheral blood of six donors, sampled three times over the course of a year. We tracked the most abundant B cell clonal lineages in time and analyzed their cell subset and isotype composition, phylogenetic history, and mode of selection.

In all individuals, the observed clonal lineages clearly fell into two clusters. HBmem represents persistent memory with a predominant IgM isotype; such clonal lineages were equally sampled from all time-points and rarely included ASC clonotypes. The MRCA of observed clonotypes in HBmem lineages almost matched the predicted germline sequence — and in 14.5% of the lineages, matched completely — indicating that the probability of observing a clonotype from these lineages has no association with the position in that lineage's phylogeny. Horns and colleagues observed lineages with very similar features to HBmem, which also possessed persistent dynamics against a background of vaccine-responsive lineages and were predominantly composed of the IgM isotype (Horns et al. 2019). However, their study was performed on bulk B cells, so there was no possibility to track their relatedness to the Bmem subset.

In contrast, the LBmem cluster demonstrates completely different features, with lineages largely composed of ASC clonotypes with switched IgA or IgG isotypes, showing active involvement in ongoing immune response. The MRCA of LBmem lineages differed from the germline sequence by some number of SHMs, and only 1.9% of LBmem lineages had a complete match between the MRCA and the germline sequence. A large G-MRCA distance implies that the observed clonotypes originated from an already-hypermutated ancestor, and that we had therefore sampled clonotypes from a single clade of the lineage phylogeny. Such an effect can be caused both by rapid expansion of the clade and migration of the clade's clonotypes, diverged in the tissue of residence (Mandric et al. 2020). We also observed that most LBmem lineages expanded at T2 or T3 (38 out of 45, > 80%) had at least one clonotype detected in the Bmem subset at the previous time-point, leading us to conclude that LBmems represent the progeny of reactivated memory B cells. We found one clear example which further supports this idea: a lineage that possesses all features of the HBmem cluster except for one monophyletic clade, typical for LBmem lineage. This example of HBmem-LBmem transition is very similar to reactivated persistent memory, as observed by Hoehn *et al.* in response to seasonal flu vaccination (Hoehn et al. 2021). In addition, Phad *et al*. have recently demonstrated clonal relatedness of the emerging plasmablasts to the persistent Bmem lineages in longitudinal immune repertoire profiling in aged healthy donors (Phad et al. 2022). Thus, it can be assumed that at least part of the observed LBmem lineages represent the progeny of the persistent memory represented by HBmem lineages.

Our analysis of the selection mode in HBmem and LBmem lineages supported our assumptions. We showed that both lineages experienced positive selection from the germline sequence to the MRCA of the observed clonotypes — as expected, assuming that primary B cell activation is followed by affinity maturation associated with clonal lineage expansion. However, the pressure of positive selection was stronger in LBmem lineages than in HBmem. In addition, we detected an excess of sites under positive selection in LBmem lineages that underwent evolution after the MRCA as well. This leads us to the hypothesis that LBmem cells underwent additional rounds of affinity maturation after reactivation. *Hoehn et al.* did not study the mode of selection in their reactivated lineages, but some clonotypes were sampled from germinal centers, supporting the involvement of affinity maturation in the process of memory reactivation. In subsequent evolution after the MRCA, we detected negative selection in both types of lineages — but again, stronger in LBmem. This excessive negative selection in LBmem lineages can be considered as a signature of purification of the clonal lineage from deleterious BCR variants in the course of affinity maturation.

Obviously our study is done under certain limitations. First, we used a relatively small cohort group, which were combined from donors with different allergy status. Nevertheless, it was large enough to reveal that all our observations were reproducible among donors independent from their health conditions. We also observed no evidence that allergy status affects the structure of our data, which allowed us to generalize obtained observations for the whole cohort group. Second, we have not observed much direct evidence of the process of memory reactivation and new rounds of affinity maturations. Reactivation process was clearly detected in only one clonal lineage (**Figure 3.3F**). However, this explanation of the given data is convincing for us because of the whole set of indirect evidence, such as large G-MRCA distance and close relatedness of LBmem clonotypes, the presence of Bmem clonotypes prior to LBmem expansion and different modes of natural selection in HBmem and LBmem clusters. Our hypothesis is also supported by recent studies (Hoehn et al. 2021, Phad et al. 2022). Nevertheless, we do not deny the possibility of other mechanisms, staying behind the structure of our data, so the functioning of B cell immunity in homeostasis conditions requires the following investigation.

Thus, in this work, we performed a detailed longitudinal analysis of BCR repertoires from immune-experienced B cell subsets from donors without severe pathologies, and from these data, we have produced a framework for the comprehensive analysis of selection in BCR clonal lineages. Our results demonstrate the long-term persistence of memory-enriched clonal lineages in peripheral blood. Signs of positive selection were detected in both memory- and ASC-dominated B cell lineages. Together, the results of our evolutionary analysis of B cell clonal lineages coupled with B cell subset annotation suggest that the reactivation of pre-existing memory B cells is accompanied by new rounds of affinity maturation.

## Contribution

My contribution to this work was in the general analysis of preprocessed most abundant clonal lineages, including their composition and temporal dynamics (**Figures 3.2**, **Supplementary Figures A-1** and **A-2**), as well as ongoing analysis of lineage's phylogenies (**Figure 3.3**, **Supplementary Figure A-3**) and analysis of the selection mode, affecting lineage's evolution (**Figure 3.4**). The text presented in this chapter was written with the contribution of all coauthors of (Mikelov et al., 2022).

# CHAPTER 4: SARS-CoV-2 ESCAPE FROM CYTOTOXIC T CELLS DURING LONG-TERM COVID-19

Here we consider the complex system of intrahost interaction between pathogens and adaptive immunity from the side of pathogens on the example of SARS-CoV-2. SARS-CoV-2 immune escape from broadly-neutralizing antibodies by modification of their binding is well documented (Kemp et al. 2020; Williamson et al. 2021; Khatamzas, Rehn, et al. 2021; Starr et al. 2021; Garrett et al. 2021), however T cell escape is much less described. Indeed, intra-host escape from CD8 T cells was described for other long-term infections including HIV-1 and hepatitis C (Bronke et al. 2013; Troyer et al. 2009; Goulder et al. 2001; Erdmann et al. 2015; Erickson et al. 2001). This Chapter is devoted to the analysis of intrahost viral evolution during long-term COVID-19 under condition of depleted B cell immunity without the treatment by convalescent plasma. We hypothesized that such immune conditions could force the viral population to avoid T cell recognition.

## Methods

### Sample collection and sequencing

Special informed consent was obtained from the patient before the specimen for additional tests were taken. RT-PCR of swabs and sequencing of viral RNA was performed in the Smorodintsev Influenza Research Institute. All specimens were obtained and transported according to standard sampling protocol. RNA from nasopharyngeal swabs was extracted using QIAamp Viral RNA Mini Kit (QIAGEN). RNA from patient A post-mortem FFPE specimens was extracted using RNeasy FFPE Kit (QIAGEN). Samples were tested for SARS-CoV-2 viral RNA by real-time RT-PCR on thermal cycler CFX96 (BioRad) using "Intifica SARS-CoV-2" Kit (Alkor Bio). Whole-genome amplification of SARS-CoV-2 virus genome for samples from August 2020 and from January 2021 was performed using BioMaster RT-PCR Premium kit (Biolabmix) and primers from ARTIC Network protocol version 3(Tyson et al. 2020) and ARTIC Network protocol version 1 (Itokawa et al. 2020) with modifications, respectively. Nextera XT (Illumina) kit was used for library preparation in August 2020 and DNA Prep (Illumina) kit was used for library preparation in January 2021, and the libraries were sequenced using the MiSeq platform (Illumina) with version 3 600-cycle chemistry.

The DNA of patient S was extracted from peripheral blood using QIAmp Blood DNA Mini kit. DNA sample was prepared and captured with the SureSelect Human All Exon kit v7 (Agilent), and whole exome was sequenced using MGISEQ-2000 at Pirogov Russian National Research Medical University (Moscow, Russia).

### Flow cytometry assays

Flow cytometry assays were performed using cryopreserved PBMCs. Cells were isolated from patients' heparinized blood by gradient centrifugation with lymphocyte separation medium Lymphosep (BioWest), frozen in freezing medium containing 10% DMSO (AppliChem) in FBS (Gibco) and stored in liquid nitrogen until usage.

For B-cells analysis presented in **Supplementary Figure B-5**, PBMCs samples were towed in a 37ºC water bath and stained with fluorescently-labeled antibodies to surface markers CD19-APC/Fire 750 (Clone: SJ25C1, Biolegend), BV421-CD20 (Clone: 2H7, Biolegend), CD3-BV605 (Clone: OKT3, Biolegend). PBMCs from a healthy volunteer were used as a control. B-cells were identified as a live CD3 $^{-}$/CD19 $^{+}$/CD20 $^{+}$ population.

The T-cell response was assessed by intracellular cytokine staining. For further analysis, cells were towed in a 37ºC water bath and stimulated for 6 hours with 5 µg/ml of the commercial available peptide mixture of SARS-CoV-2 proteins S, N, M, ORF3a and ORF7a (Generium, Russia) (for **Supplementary Figure B-6B:C**) or one of the peptides PTDNYITTY, PADNYITTY or PPDNYITTY or peptide pools (YLQPRTFLL + STNVTIATY + KPRSQMEIDF + GPQNQRNAPRITF + VPLHGTIL and YLQPSTFLL + SINVTIATY + KLRSQMEIDF + GTQNQRNAPRITF + VPLHGTIR) (for **Figure 4.3**) in the RPMI medium (Gibco), containing 10% of FBS (Gibco), 1% of penicillin-streptomycin solution (Gibco), Brefeldin A (BD) and costimulatory CD28/CD49 reagent (BD). Negative control samples were stimulated with the complete medium; for positive control, PMA/ionomycin (Sigma) combination was used. Surface markers were stained with fluorescent antibody panel containing CD3-APC/Fire (Clone: SK7, Biolegend), CD4-AF647 (Clone: SK3, Biolegend), CD8a-AF600 (Clone: HIT8a, Biolegend), CD45RA-PE/Dazzle (Clone: HI100, Biolegend), CD197-BV421 (Clone: G043H7, Biolegend). Intracellular cytokines were stained using IL-2-FITC (Clone: MQ1-17H12, Biolegend), IFNγ-PE (Clone: 45.15, Beckman Coulter), TNFα-BV785 (Clone:

MAb11, Biolegend) antibodies. Cells were permeabilized with BD Cytofix/Cytoperm™ Fixation/Permeabilization Solution Kit (BD) according to the manufacturer's instructions. Data were collected on a CytoFlex flow cytometer (Beckman Coulter). The results were analyzed using the Kaluza Analysis v2.1 program (Beckman Coulter). Interleukin (IL) 2, interferon γ (IFNγ) and tumor necrosis factor (TNFα) response was measured in effector memory T cells (Tem). To identify Tem, lymphocytes were gated based on their size and granularity. Live CD3$^{+}$T cells were identified and subdivided into CD4+ and CD8+ T cells. These populations were further subdivided based on the expression of CD45RA and CD197(CCR7). CD3+ CD4+ or CD3+ CD8+ lymphocytes with the CD45RA-/CCR7- phenotype were considered Tem cells (**Supplementary Figure B-6A**). Cut-off values for the definition of cytokine-producing T cell responses stimulated with SARS-CoV-2 peptides were ≥5 events and a ≥2-fold difference in the magnitude of TNF $^{+}$, IFNγ $^{+}$ or IL-2 $^{+}$ Tem cells compared to the non stimulated control.

### Virus isolation and antigenicity

Live viruses (samples 30579V and 30769V from August 20, 2020 and 22748V and 23680V from February 19, 2020) were isolated from patient S swab samples in Vero E6 cells (IZSLER #BSCL87). Culture was inoculated for 2 hours with swab material diluted 1/10 in DMEM (Biolot) supplemented with 2% HI-FBS (Gibco), 1% anti-anti (Gibco) and then incubated for 3 days until first CPE signs. Samples were subsequently passaged one time in Vero cells (ATCC #CCL81).

A total of 16 serum samples were obtained during the first wave of the COVID-19 pandemic in spring-summer 2020 from recovered volunteers with PCR-confirmed SARS-CoV-2 infection and tested in a microneutralization assay.

Microneutralization was performed with hCoV-19/St_Petersburg-3524V/2020 virus (GISAID EPI_ISL_415710, with the ΔF combination of mutations absent, designated as Reference), and 30769V and 23680V viruses isolated from the patient S (designated patient S August 2020 and patient S January 2021, respectively). Serum was heat inactivated (56°C, 60 min), serially diluted 2-fold starting from 1/10, mixed with 25 TCID50 of virus, incubated for 1h at 37°C and inoculated into Vero cells in triplicates in 96-well plate. 5 days after inoculation, neutralizing antibody titer was calculated as the reciprocal of the highest serum dilution preventing CPE.

Serum samples obtained from patient S were tested for virus specific antibodies in ELISA and in microneutralization assay with either Reference or patient S viruses. ELISA was performed with "SARS-CoV-2-IgG-IFA-BEST" (VEKTOR BEST #D-5501) according to the manufacturer's instructions.

### HLA genotyping

HLA genotyping was performed using a commercial kit according to the manufacturer’s instructions (PARallele™ HLA solution v3, Parseq Lab). HLA-A, -B, -C, -DRB1 and -DQB1 loci were genotyped with 3-field resolution. Simultaneously, HLA calling was performed from WES data using HLA-HD version 1.3.0 (Kawaguchi and Matsuda 2020) with IPD-IMGT/HLA database Release Version 3.43. The inferred alleles are listed in **Supplementary Table B-3**.

Using HLA-2-Haplo software tool (Geffard et al. 2020) this set of alleles was split into two haplotypes presented in **Supplementary Table B-3**. A European population database was used in this procedure. An a-posteriori probability of found combination was 97.6%. As one can see, the found haplotypes are among the most common variants in the European population.

### Consensus calling

Raw reads were trimmed with Trimmomatic version 0.39 (Bolger, Lohse, and Usadel 2014) to remove adapter sequences and low-quality ends. Trimmed reads were mapped onto the Wuhan-Hu-1 (MN908947.3) reference genome with BWA MEM version 0.7.17 (Heng Li 2013). The following reads were then removed from the mapping: reads with abnormal insert length to read ratio (greater than 10 or less than 0.8), reads with insert length greater than 1100, reads with more than 50% soft-clipped bases. Soft-clipped ends were trimmed from the remaining reads, 10 nucleotides were cropped from read ends using custom scripts, and primer sequences were removed with ivar version 1.3 (Grubaugh et al. 2019). Only reads with at least 30 nucleotides remaining after the procedure were kept.  SNV and short indel calling was done with LoFreq version 2.1.5 (Wilm et al. 2012), with SNVs considered consensus if they were covered by at least 4 reads and supported by more than 50% of those reads; indels were considered consensus if they were covered by at least 20 reads with at least 50% of those supporting the variant. Regions that were covered by fewer than 4 reads were masked as NC. We attributed several positions that were covered by 2 or 3 reads, but matched the reference and were conserved throughout all samples (22612, 23680, 24160, 27064, 30579 and 30769), to the reference; as these positions did not mutate, this decision did not affect any of our analyses. Consensus was created by bcftools version 1.9 (H. Li 2011; Heng Li et al. 2009) consensus.

### Phylogenetic analysis

255,389 genomes of SARS-CoV-2 were downloaded from GISAID on December 12, 2020, and aligned with MAFFT v7.453 (Katoh and Standley 2013) against the reference genome Wuhan-Hu-1/2019 (NCBI ID: MN908947.3 (Wu et al. 2020) with --addfragments --keeplength options. 100 nucleotides from the beginning and from the end of the alignment were trimmed. After that, we excluded sequences (1) with more than 300 positions of missing data (Ns) and gaps, (2) excluded by Nextstrain (https://github.com/nextstrain/ncov/blob/master/defaults/exclude.txt), or (3) from non-human animals other than minks, leaving us with 201,948 sequences. Identical sequences were then collapsed within the country and host and annotated by the Pangolin package version 2.1.0 (Rambaut et al. 2020). To this dataset, we added the two patient S samples obtained in August, 2020 as well as the patient A sample. As sequences of patient S belonged to the B.1.1 lineage, we further only kept sequences annotated as B.1.1, excluding a large clade defined by mutation G25563T (GH clade in GISAID (Sergei Pond 2020) nomenclature). For the purposes of phylogenetic analysis, we additionally masked the highly homoplasic site 11083. The final set of 49,083 sequences was used to construct the phylogenetic tree with IQ-Tree v2.1.1 (Nguyen et al. 2015) under the GTR substitution model and '-fast' option. Ancestral sequences at the internal tree nodes were reconstructed with TreeTime v. 0.8.0 (Sagulenko, Puller, and Neher 2018). Having ensured that the two patient S samples form a clade rooted at the patient A sample and not carrying any samples other than those of patient S, we then separately reconstructed the phylogeny of all six samples of patient S, rooted it with patient A, and manually added the resulting clade to the downsampled B.1.1 tree. For visualization purposes, the tree was downsampled to contain 1% of samples, including the patient A sample and the complete clade containing all patient S samples.

To estimate the molecular clock rate of the patient S lineage (**Figure 4.1C**), we downloaded all sequences available in GISAID on May 31, 2021, filtered them as described above, and subsampled the filtered dataset to 50,000 samples preserving all Russian sequences. To this dataset, we added the six patient S samples and the ancestral patient A sample. We then aligned the obtained 50,007 sequences against the reference sequence and reconstructed the phylogeny with Fasttree version 2.1.11 (Price, Dehal, and Arkin 2010). Finally, we computed root-to-tip distances and calculated the slope of the root-to-tip distance vs. sampling dates regression line for the three separate datasets: (1) patient S samples, (2) B.1.1.7 samples, and (3) the remaining samples from the subsampled GISAID dataset. To validate the difference between the estimated clock rates for patient S samples and samples belonging to dataset (3), we subsampled this dataset, picking six random samples collected on the same dates as the patient S samples, and computed the linear regression slope, in each of the 10,000 trials. (For dataset (2), this procedure was impossible because there were no B.1.1.7 samples in August 2020). None of the 10,000 samples resulted in the estimated clock rate above $15.3*10^{-4}$, implying the p-value of <0.0001.

### Effect of viral mutations on antigen presentation

To study the effect of mutations in SARS-CoV-2 proteins on their antigen presentation, we adapted a pipeline from Marty and et. (Marty et al. 2017) (**Figure 4.2A**). For each mutated site in both its ancestral and derived states, we inferred all possible peptides of certain lengths overlapping it, and calculated their percentile ranks (Rank_El) relative to a set of random natural peptides by netMHCpan version 4.1 and netMHCIIpan version 4.0 (Reynisson et al. 2020) for HLA I and HLA II respectively. We used peptide lengths between 8 and 12 amino acids for HLA I alleles, and between 12 and 18 amino acids for HLA II. If the mutated site was not presented by any of the HLA alleles either in the ancestral or derived states, we excluded it from analysis. To exclude non-presenting peptides, we used the percentile rank < 2% threshold for HLA I, and < 10% threshold for HLA II, as recommended by the netMHCpan manual. For derived states of deletions, we extended the peptide in the C-direction as necessary to preserve its length. We paired the predicted A and B chains of HLA class II alleles as suggested in the tool allele list: HLA-DQA10101-DQB10501, HLA-DQA10501-DQB10201, HLA-DPA10103-DPB10402, HLA-DPA10103-DPB10401, DRB1_0301, DRB1_0101. We excluded the stop-codon producing mutation ORF8:Q18* from comparisons of ancestral and derived states, since the corresponding values for the derived state were undefined.

As in Marty et al. (Marty et al. 2017), we used the best percentile rank (BR) among all possible peptides overlapping the mutated site as the presentation score of this site for the particular HLA allele. To estimate the overall presentation of the site in the patient, we calculated the patient harmonic best rank (PHBR), i.e., the harmonic mean of BRs of HLA alleles of the same class. To compare the effect of a mutation on site presentation, we calculated the fold change of PHBR score as the ratio of the derived PHBR to the ancestral PHBR (so that fold change > 1 indicates weakening of presentation).

To focus on the peptides shown to be immunogenic to T cells in other SARS-CoV-2 infected patients carrying the same HLA alleles as patient S, we used IEDB (Vita et al. 2019) (Immune Epitope Database and Analysis Resource, accessed on June 1, 2021) with the "positive assay only" filter. For those sites inferred to be contained in immunogenic peptide, we calculated the best percentile rank of immunogenic peptide overlapping the site of mutation (imBR).

**Population-level effects of mutations**

To check the effect of detected SARS-CoV-2 mutations on presentation by the HLA alleles other than those of patient S, we calculated the BR scores as explained above for the most frequent classical HLA alleles of each family that together represented 95% of the HLA alleles in the world population. The list and frequencies of such alleles were taken from Sarkisova et al. and Solberg et al. (Solberg et al. 2008; Sarkizova et al. 2020).

For most mutations detected in immunogenic epitopes, at least one of the HLA I alleles of patient S demonstrated extreme values of BR fold change in comparison with other alleles (**Figure 4.4A**). To check the probability of such an observation happening by chance, we performed a permutation test, calculating the probability that a randomly chosen set of alleles has the same or a more extreme value of mean BR fold change across all mutations overlapped by immunogenic peptides as that of alleles of patient S. This was true for 33 out of 100000 permutations, corresponding to $p = 0.0033$ (**Figure 4.4B**). None of the HLA II immunogenic epitopes overlapped any of the mutated sites; the only mutated site adjacent to such an epitope (S:S50L) did not stand out in the permutation test ($p = 0.6996$; **Figure 4.4C:D**).

To compare the effects of mutations between different HLA alleles in **Figure 4.4E:F**, we calculated the mean BR across all changed sites. This analysis again excluded ORF8:Q18*, which nevertheless prevented production of high-affinity epitopes for most alleles.

### Data analysis and visualization

Data analysis was performed in R version 4.0.0 (R Core Team 2018), and figures were visualized with ggplot2 package version 3.3.2 (Ginestet 2011). SARS-CoV-2 phylogenetic tree was visualized with ITOL version 6 (Letunic and Bork 2007).

### Data availability

Sequence data is available from the Sequence Read Archive: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA749008/ (SRA: PRJNA749008). Consensus sequences are available from the GISAID.

### Code availability

Code is available at https://github.com/EvgeniiaAlekseeva/patient_S.

### Ethics declaration

The study was approved by the Local Ethics Review Board of the Smorodintsev Research Institute of Influenza and by the Biomedical Ethics Committee of the I.P. Pavlov First Saint Petersburg State Medical University. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

## Results

### Case description

Patient S (**Supplementary Note B-1**), a female previously diagnosed with Non-Hodgkin's diffuse B-cell lymphoma IV stage B, tested positive for SARS-CoV-2 for the first time on April 17, 2020. In the preceding week, she had had close contact with patient A, who later died of COVID-19 pneumonia; paraffin blocks with post-mortem material of patient A were subsequently analyzed for SARS-CoV-2 by PCR, followed by RNA extraction and sequencing, as a probable source of infection. Patient S has undergone three periods of positive tests, alternating with two periods of negative tests, between April 17, 2020 and March 1, 2021, spanning a total of 318 days (see **Figure 4.1A**, **Supplementary Table B-1**). She had symptoms of severe COVID-19 between June 6 - September 1, 2020 (**Supplementary Figure B-1A:B**), and again between January 9 - March 1, 2021 (**Supplementary Figure B-1C**), including subfebrile fever and pneumonia with typical COVID-19 patterns. We isolated live viruses from swab samples obtained in both of these periods (August 20, 2020 and February 19, 2021).

Between April 30, 2020 and February 16, 2021, patient S received several cycles of chemotherapy under several different regimens, including monoclonal antibody rituximab. Courses of chemotherapy were typically followed by a decrease in white blood cell counts, especially lymphocyte and neutrophil counts, to values below the normal range (**Supplementary Figure B-2**). On December 28, 2020, autologous haematopoietic stem cell transplantation (auto-HSCT) was performed. In January 2021, near the end of the study period, patient S received three doses of convalescent plasma. Six nasopharyngeal swab samples suitable for next generation sequencing, together spanning 308 days of the disease, were obtained, alongside two blood samples (**Supplementary Table B-1**).

### Intrahost evolution of SARS-CoV-2

Whole-genome sequencing was performed for six nasopharyngeal swab samples obtained from patient S in August 2020 - February 2021, as well as for an April 2020 sample obtained from patient A (**Figure 4.1A**). Phylogenetic analysis (**Supplementary Note B-2**) indicates that both PCR positive periods of patient S in August 2020 and January-February 2021 constitute a single infection. Indeed, all patient S samples formed a single clade within the B.1.1 lineage on the global SARS-CoV-2 phylogeny, with the patient A sample as its ancestor (**Figure 4.1B**). No other Russian samples available in GISAID nest within the patient S clade (**Figure 4.1B**), indicating that the virus evolved in patient S has not seeded observable onward transmission.

The two August 2020 samples were characterized respectively by 12 and 18 mutations specific to patient S. In turn, the January-February 2021 samples gained additional 10 to 21 changes. Overall, a total of 40 changes compared to the ancestral state were observed in at least one of the samples, 34 of which were observed by the end of the study period (**Supplementary Note B-3**). This corresponds to the point substitution rate of $15.3\text{x}10^{-4}$ per nucleotide per year, which substantially exceeds the evolutionary rate of SARS-CoV-2 in the general population (permutation test, $p<10^{-4}$; **Figure 4.1C**). Nearly all accumulated changes were detected in samples obtained before convalescent plasma transfusions (**Figure 4.1A,D; Supplementary Table B-1**), indicating that these transfusions could not have affected the observed viral evolution.

The accumulated mutations occurred throughout the viral genome, affecting 18 of the 26 viral genes (**Figure 4.1D**). However, there was an excess of nonsynonymous changes in the genes encoding surface proteins: out of the 25 changes, 8 (32%) fell in the spike gene which by length constitutes 13% of the viral genome, while 2 (8%) fell in the envelope gene which constitutes 0.8% of the genome (two-sided Binomial test, p = 0.018 and 0.016, respectively). Many of the observed amino acid substitutions were indicative of positive selection in the general population (**Supplementary Note B-4**), and some were previously implicated in antibody escape (**Supplementary Note B-4**). However, virus evolution did not lead to detectable reduction in sensitivity to neutralizing antibodies by the end of the study period compared to a prototype viral strain (**Supplementary Figure B-4**).

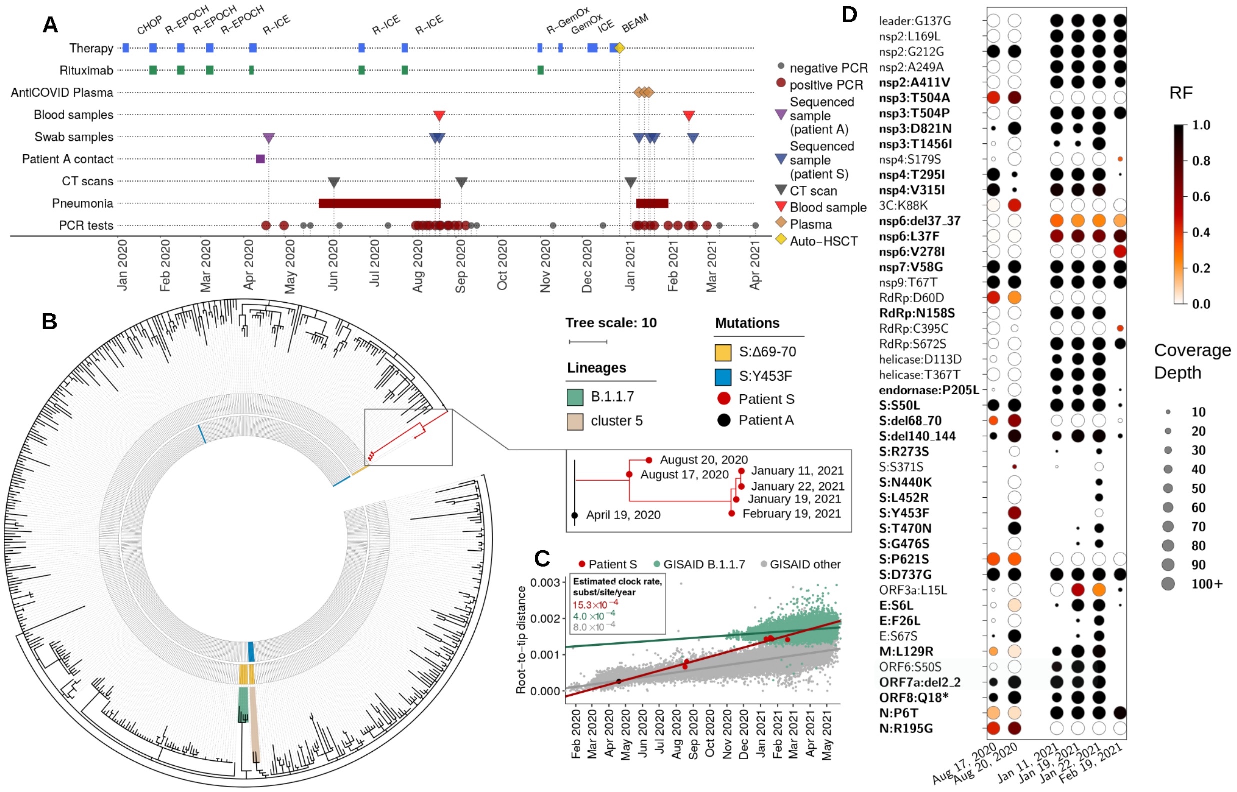


**Figure 4.1. Intrahost evolution of SARS-CoV-2 in patient S. A:** The timeline of patient S disease and therapy. **B:** The phylogenetic tree of B.1.1 pruned to contain a random set of 1% of all samples, including the patient A sample (black dot) and the complete clade carrying the patient S samples (red dots). The 2020 samples carried the ΔF combination of mutations (S:Δ69-70HV and S:Y435F; Supplementary Note B-3) marked in the two inner circles in yellow and blue. The B.1.1.7 lineage and cluster 5 are shaded. **C**: Regression of root-to-tip genetic distances vs. sampling dates, for patient S samples (together with the ancestral sample of Patient A), B.1.1.7 lineage GISAID samples, and other GISAID samples. Estimated slopes (molecular clock rates) are provided in the inset. In **B** and **C**, the consensus nucleotides (i.e., those supported by more than 50% of the reads, RF>50%) were used to position patient S and A samples. **D**: Variant frequencies in the six patient S samples. All consensus variants (RF>50%, N=40) and non consensus variants with 30%<RF<50% (N=7) are shown (Supplementary Table B-2). The figure is adapted from (Stanevich et al., 2023).

### Host immune response

To understand the functional features of immune response in patient S, we analyzed her blood samples collected at multiple timepoints spanning the course of the disease (see Methods, **Supplementary Table B-1**). Flow cytometry revealed the absence of B lymphocytes throughout the period of PCR positivity (**Supplementary Figure B-5**). Blood serum samples were also analyzed by ELISA for IgG antibodies specific to the SARS-CoV-2 S-antigen; a weak IgG response was registered in one of the samples but no response in the remaining samples. No neutralizing antibodies were detected at any time point by a VN assay using live SARS-CoV-2 strain (**Supplementary Table B-1**).

By contrast, we detected a pronounced SARS-CoV-2 specific T-cell response. Indeed, *in vitro* stimulation with a peptide mixture of SARS-CoV-2 proteins (S, N, M, ORF3a and ORF7a) caused an expansion of SARS-CoV-2-specific CD4 and CD8 effector memory T-cells (Tem) at both time points (**Supplementary Figure B-6**).

### Mutational escape from cytotoxic T cells

Given the absence of B-cell but the presence of T-cell immune response in patient S, we hypothesized that the 31 amino acid sequence-altering mutations acquired by SARS-CoV-2 may have led to escape from T cell immunity. First, we asked if these mutations affect presentation of the peptides carrying them by the HLA alleles of patient S (**Supplementary Table B-3**). For this, we adapted an existing pipeline(Marty et al. 2017) to calculate the PHBR (patient harmonic best rank) score (**Figure 4.2A**) for both the ancestral and the derived state at site of each of the 30 mutations (except ORF8:Q18*, **Supplementary Note B-5**). Most sites could be presented in their ancestral state by at least one HLA allele of both classes (27 out of 30 by HLA I, and 24 out of 30 by HLA II). We found that five of the observed mutations substantially (>3-fold) increased the PHBR score for the peptides presented by HLA I, indicating impaired presentation (**Figure 4.2B**). One of these mutations, S:del141-144, also increased the PHBR score for HLA II (**Figure 4.2C**).

While an increase in PHBR score can help a peptide escape antigen presentation, this can only affect T cell response if the corresponding peptide is recognised by T cells. To specifically address the effect of mutations on immunogenic peptides, we used IEDB (Vita et al. 2019) to obtain the list of SARS-CoV-2 peptides that were shown to be immunogenic in complexes with the HLA alleles carried by patient S. There were 17 such peptides for HLA I alleles, together overlapping the sites of 11 of the mutations (some of the sites were covered by more than one peptide) (**Supplementary Table B-4**). All these mutations were fixed in the course of intra-host evolution by the end of the study period. No HLA class II immunogenic peptides covering the changed sites were found in IEDB. To focus on the immunogenic peptides, we calculated the imBR (immunogenic best rank) for each of these sites in the ancestral state and compared it to the corresponding value for the derived state. The mutations strongly decreased presentation of immunogenic peptides, indicating that they cause escape from CD8 T cell response (**Figure 4.2D**). Together with ORF8:Q18* which prevented presentation of the bulk of the ORF8 protein (**Supplementary Note B-5**), this totals to 12 changes with cytotoxic T cell escape effect.

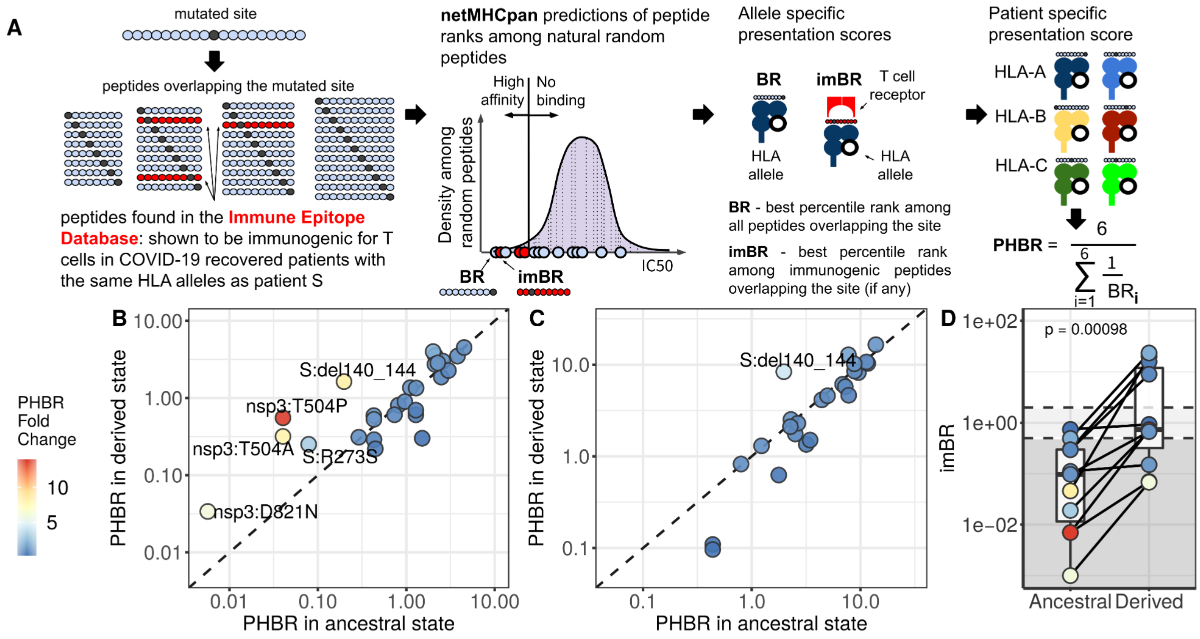


**Figure 4.2. Mutational escape from cytotoxic T cells. A**: Calculation of site presentation scores (adapted from Marty et al. (Marty et al. 2017)). **B, C**: Change of PHBR scores caused by mutations for HLA I (**B**) and HLA II (**C**) respectively. Dot color corresponds to PHBR fold change; the mutations that substantially (>3-fold) increase PHBR are signed. Sites that did not bind any of the patient's HLA alleles both in ancestral and derived states are not shown. **D**: Comparison of imBR scores for the mutated sites in their ancestral and derived states. The level of significance is calculated by the Wilcoxon sign-rank test.

### Tracking the viral escape

Next, we assessed the change in T-cell response caused by the observed mutations. First, we focused on the two mutations causing the largest PHBR fold change (**Figure 4.2B**). These were the two mutations at position 504 of the nsp3 protein, nsp3:T504A and nsp3:T504P, which were fixed sequentially at the first (T1, August 20, 2020) and the second (T2, February 19, 2021) sampled time points respectively (**Figure 4.1D**). We asked how well the peptides covering these three amino acid variants elicited T-cell response in samples corresponding to these time points. We used the highest ranking peptides covering the mutated site in its ancestral (P**T**DNYITTY) and derived (P**A**DNYITTY, P**P**DNYITTY) states; P**T**DNYITTY was previously shown to be immunogenic in complex with the HLA-A:01*01 allele which is carried by patient S (Ferretti et al. 2020; Gangaev et al. 2021; Saini et al. 2021).

In the T1 sample, when just nsp3:T504A was detected at intermediate frequencies (**Figure 4.1D**), *in vitro* stimulation of CD8+ T cells indicated response to both the ancestral (P**T**DNYITTY) and the derived (P**A**DNYITTY) peptide changed by nsp3:T504A (**Figure 4.3**). This response was mediated primarily by polyfunctional IFNγ$^+$IL2$^-$TNFα$^+$ effector memory T-cells. The response to P**A**DNYITTY was slightly weaker (0.035% vs 0.043% of effector CD8 T cells), suggesting a partial escape caused by nsp3:T504A. Stimulation by P**P**DNYITTY corresponding to the nsp3:T504P allele caused no cytokine response in the T1 sample. In the T2 sample (**Figure 4.1A**), when nsp3:T504P was already fixed, still no cytokine response to P**P**DNYITTY was observed, confirming invisibility of this peptide to cellular immune response due to weak binding with HLA. Response to P**T**DNYITTY and P**A**DNYITTY also vanished at T2; this could indicate that the CD8 T cell clones specific to T and A amino acids became irrelevant with the loss of the corresponding viral variants, and got no antigenic re-stimulation that could drive clonal expansion after auto-HSCT (Mamedov et al. 2011).

Next, we explored the T-cell response to the pool of peptides corresponding to virus epitopes that gained amino acid mutations between August 2020 and January 2021. The pool included 5 peptides with previously confirmed immunogenicity and characterized by a strong change in PHBR due to the observed mutations, indicating a probable escape from the HLA alleles of patient S. We compared the pool of peptides in their ancestral state: YLQP**R**TFLL (S:R273S), S**T**NVTIATY (nsp3:T1456I), K**P**RSQMEIDF (endornase:P205L), G**P**QNQRNAPRITF (N:P6T) and VPLHGTI**L** (M:L129R), to the corresponding peptides with acquired amino acid mutations that resulted in weak or no binding: YLQP**S**TFLL, S**I**NVTIATY, K**L**RSQMEIDF, G**T**QNQRNAPRITF and VPLHGTI**R** (**Supplementary Table B-4**).

At time point T1, we found a pronounced subpopulation of polyfunctional (IFNγ+/TNFα+) cytokine producing CD8 T cells responding to initial non-mutated peptides. This subpopulation comprised 0.95% of effector CD8 T cells, indicating a strong T cell response to this set of epitopes (**Figure 4.3C**). Meanwhile, no T cell response was observed against the pool with acquired mutations, confirming immunoediting-driven origin of the observed mutations. Both peptide pools showed negligible response at time point T2, presumably due to poor post-HSCT expansion of T cell clones in the absence of the cognate antigenic stimulus. Prior to the escape, the CD8 T-cells responding to the pool of the 5 escaping peptides (0.95%), together with the peptide changed by nsp3:T504P (0.045%), constituted as much as ~1% of the total effector CD8 subset, and this response has been fully eliminated by viral escape.

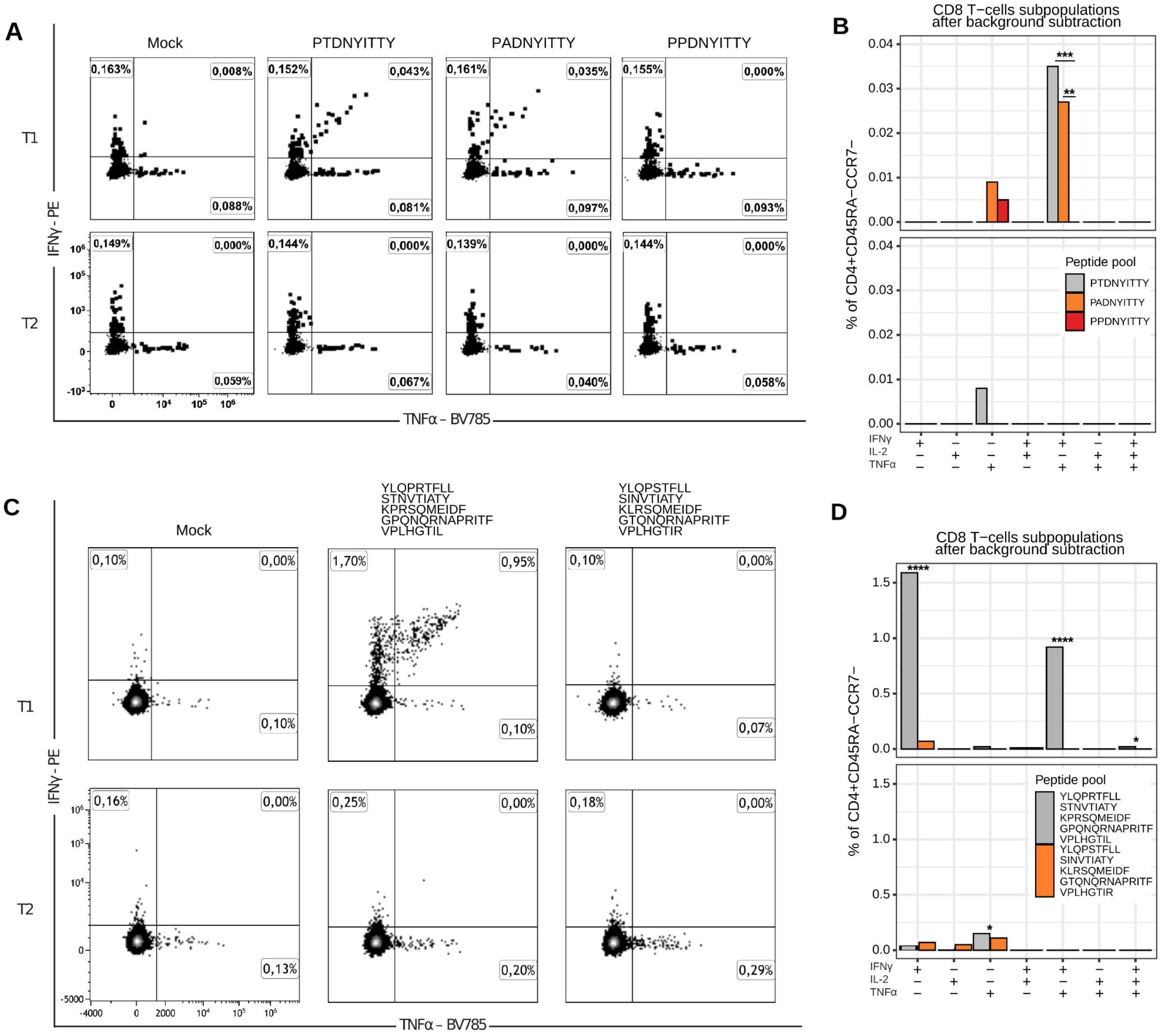


**Figure 4.3. The CD8 T cell immune response to the SARS-CoV-2 epitopes in ancestral and derived states. A, C**: Flow cytometry plots showing the cytokine profiles of SARS-CoV-2-specific CD8 effector memory T cells after stimulation with epitopes carrying the ancestral and the two derived (nsp3:T504A and nsp3:T504P) amino acid variants (**A**), and pools of 5 immunogenic HLA binders before and after acquiring the binder-escape mutations (**C**). Amino acid variants corresponding to ancestral and derived states highlighted by gray and red colors respectively. **B, D**: Corresponding bar plots representing the percentage of different

cytokine-producing populations of SARS-CoV-2-specific CD8 T cells after mock background subtraction. T1, August 2020 sample; T2, February 2021 sample. Statistical significance is calculated by two-proportions Z test: * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 10^{-3}$, **** = $p \leq 10^{-4}$. The figure is adapted from (Stanevich et al., 2023).

**Possible population-level effects**

It has been suggested that escape from humoral immunity in immunosuppressed patients may give rise to SARS-CoV-2 strains with increased fitness in the general population(Harvey et al. 2021). Similarly, escape from cellular immunity in the course of intra-host evolution could affect immune response to descendant SARS-CoV-2 strains outside the host where it evolved. We aimed to estimate the possible effect of the viral evolution in patient S for the human population at large. For this, we compared the BR (**Figure 4.2A**) fold change caused by the mutations observed in patient S for the globally most frequent HLA alleles of each family that together cover 95% of worldwide population frequency (Solberg et al. 2008; Sarkizova et al. 2020). This set of alleles includes all 12 HLA alleles of both classes (I and II) of patient S, which happen to be quite frequent globally (**Supplementary Table B-3**) .

As expected, the mutations observed in immunogenic epitopes tended to escape the HLA I alleles of patient S to a larger extent than other frequent HLA I alleles (**Figure 4.4A:B**); no such difference was observed for HLA II alleles (**Figure 4.4C:D**). Nevertheless, these same mutations also reduced binding for other globally frequent HLA I alleles (mean BR fold change = 1.59, **Figure 4.4E**), although not HLA II alleles (mean fold change = 1.02, **Figure 4.4F**). This indicates that the within-host evolution in patient S indeed could facilitate escape from cytotoxic T cells in the global population.

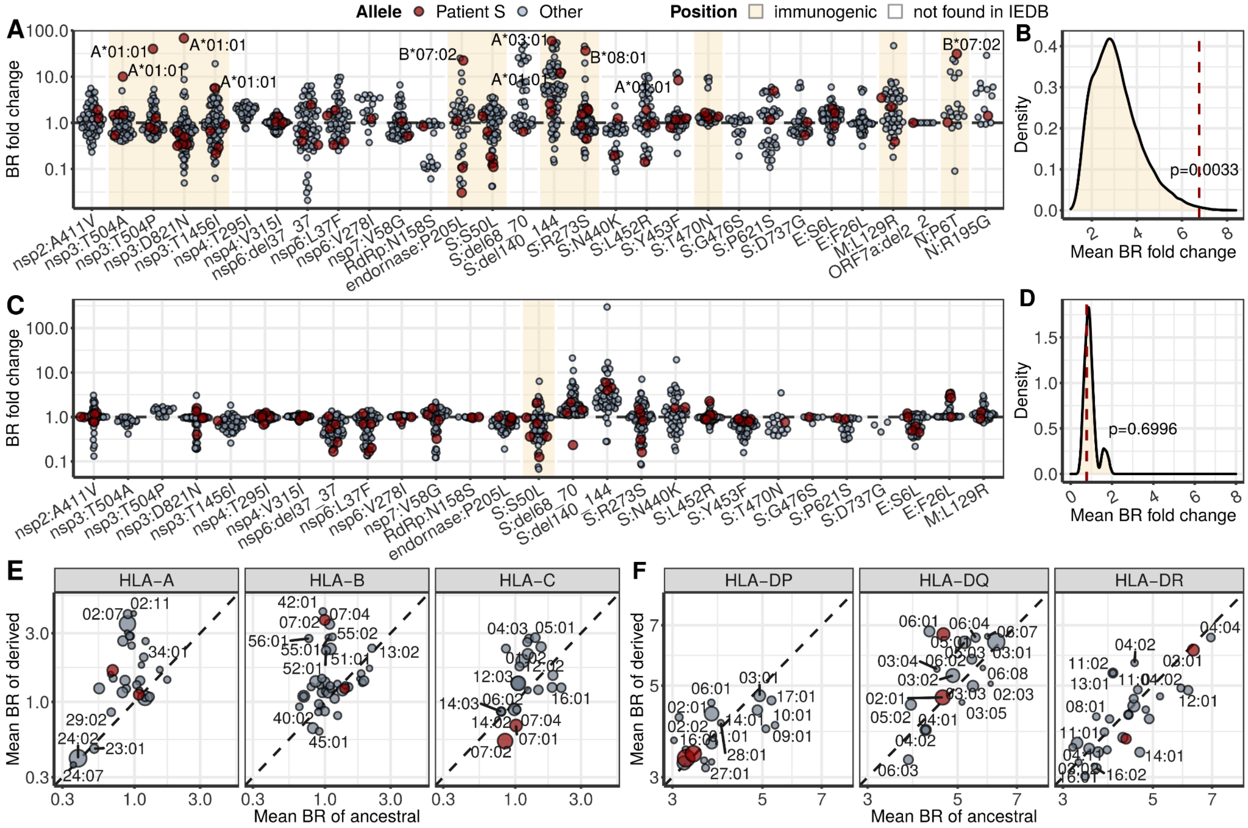


**Figure 4.4. Population-level effect of T cell escape mutations. A, C:** The effect of each of the 30 mutations observed in SARS-CoV-2 of patient S on T cell immune escape, for each of the HLA I (**A**) or HLA II (**C**) alleles carried by patient A (red) and frequent globally (gray). The mutations that change immunogenic peptides (for HLA I) or are adjacent to such peptides (for HLA II) according to IEDB are highlighted. Alleles that do not present the corresponding position in both ancestral and derived state are not shown. For the mutations that correspond to >5-fold increase in BR, the corresponding HLA alleles are signed. **B, D:** Distribution of mean BPR fold changes among immunogenic positions for HLA I (**B**) or II (**D**) alleles, based on $10^5$ random generations of individual allele composition; the red dashed line is the percentile corresponding to the allele composition of patient S. **E, F:** The sum effect of the amino acid changing mutations observed in SARS-CoV-2 of patient S on antigen presentation by the globally most frequent HLA class I (**E**) and class II (**F**). Alleles of patient S are in red.

## Discussion

We have described a case of unprecedentedly long COVID-19 characterized by a large amount of intrahost evolution. For over 10 months, an evolving SARS-CoV-2 lineage accumulated changes at a rate which substantially exceeded that in the general population, suggesting prevalent viral adaptation. Some of the observed changes recapitulated mutations previously observed in other immunocompromised patients and/or variants of concern (**Figure 2.1**, **Supplementary Note B-3**). This is consistent with the hypothesis that immunocompromised patients represent a hotspot of viral adaptation, causing "saltations" in the otherwise clock-like evolutionary rate of SARS-CoV-2 (Harvey et al. 2021); notably, such a jump could have happened at the origin of the B.1.1.7 ("alpha") variant which has attained global dominance in early 2021 (Harvey et al. 2021; Peacock et al. 2021).

Unlike previously described cases, however, the case described here is characterized by an unusual immune environment. The absence of own B cells, convalescent plasma therapy or monoclonal antibodies therapy during most of the study period indicates that the bulk of viral mutations have accumulated in the absence of humoral immune response. Instead, our data shows that evolution was largely driven by T cell escape. Our computational analysis revealed that many mutations changed the amino acid composition of known immunogenic CD8 T cell antigens and worsened or prevented their presentations on HLA class I alleles of the patient.

We experimentally tracked the viral escape by the 2 sequential mutations affecting the same amino acid position nsp3:T504 (nsp3:T504A and nsp3:T504P) and by the binder-escaping mutations in a pool of 5 immunogenic HLA binders (S:R273S, nsp3:T1456I, endornase:P205L, N:P6T and M:L129R). The elicited response in these two cases comprised 0.045% and 0.95% of all effector CD8 T cells, and this response has been eliminated by the escape. In a study of a cohort of 254 patients, the proportion of SARS-CoV-2 specific CD8 T cells in the overall IFN-γ expressing CD8 T cell pool rarely exceeded 1% and had the median of 0.2% (Moss 2022; Cohen et al. 2021). Thus our experiments clearly demonstrate that the proportion of the CD8+ T cell response subject to viral escape is substantial.

Our study has certain limitations. Most importantly, it uses data from a single patient. Several features of our case make similar cases rare. First, long-term viral persistence in COVID-19 is relatively rare overall. Second, poor disease outcome is common in immunocompromised patients (Kemp et al. 2020; Williamson et al. 2021; Khatamzas, Rehn, et al. 2021; Avanzato et al. 2020; Choi et al. 2020; Sepulcri et al. 2021; Nakajima et al. 2021; Moore et al. 2020; Hueso et al. 2020; L. Wei et al. 2020; Karataş et al. 2020; C. Y. Lee et al. 2021), and such a long period of viral persistence and high amount of intra-host evolution is extreme even among analogous studies. Third, our case stands out among the others by the absence of convalescent plasma treatment during most of the disease. Nevertheless, in a recent preprint by Khatamzas and colleagues (Khatamzas, Muenchhoff, et al. 2021), apparent T cell escape was observed in a patient with follicular lymphoma characterized by a shorter period of viral persistence and the usage of convalescent plasma. We consider that study an independent confirmation of our observations.

Another limitation is that our analysis only considered the effect of a subset of mutations. Our experimental measurements considered seven mutations, and disregarded other viral mutations that could have affected antigen presentation. Besides, some of the mutations may preserve antigen presentation but still alter the amino acid sequence of known CD8 T cell epitopes. Change of the epitope and prevention of recognition of the HLA-epitope complex by T cell receptors was previously described as a mechanism of immune escape (Bronke et al. 2013; Troyer et al. 2009; Dolton et al. 2021) and can also result in immune escape in our study. If additional escape is provided by these mutations, our appraisal of the proportion of the CD8+ T cell response negated by the viral escape is an underestimate.

Nevertheless, our results clearly indicate that immunoediting by cytotoxic CD8 clones is a prominent underappreciated factor in intrahost evolution of SARS-CoV-2. Similar to antibody escape, the T cell escape mutations acquired within an individual host may give rise to new epidemiologically important variants if they spill over to the general population. Notably, a recent study has revealed that CD8 T cell count is strongly associated with the level of intrahost diversity of the viral population in immunocompromised patients (C. Y. Lee et al. 2021). We predict that the changes observed in our study would also substantially affect SARS-CoV-2 antigenicity in the general population in case of onward transmission of the evolved variant. While no such transmission was detected in this case, our results emphasize an additional dimension of SARS-CoV-2 evolution which merits careful surveillance.

# Contribution

My contribution to this work was in the analysis of mutations's effects on viral antigens presentation on HLA alleles of the patient. It includes results from **Figure 4.2**, prediction of best candidate peptides for experimental design, used in **Figure 4.3**, and results of population-level effects of observed mutations from **Figure 4.2.** Phylogenetic analysis of patient's samples from **Figure 4.1** and **Supplementary Figure B-3** was done by K. Safina, E. Nabieva and G. Klink. Clinical data from **Figures 4.1A**, **Supplementary Figure B-1** and **B-2** was provided by O. Stanevich. Experimental data from **Figure 4.3**, **Supplementary Figure B-4**, **B-5** and **B-6** was obtained by A.-P. Shurygina and M. Sergeeva. The text presented in this chapter was written with the contribution of all coauthors of (Stanevich et al., 2023).

# CHAPTER 5: CONCLUSIONS

Interaction between pathogens and adaptive immune response within a host forms a complex and dynamic system, involving numerous interconnected elements. Coevolution is its essential feature, since both interacting players possess a single goal – to survive and to outcompete the second side. In this work we applied tools of evolutionary genomics and phylogenetic analysis to study this phenomenon from two different but complementary points of view. First we tracked phylogenies of B cell clonal lineages in healthy individuals, which allowed us to look at the history of B cell response on previous infections (Chapter 3). We adapted the broadly used approaches of evolutionary genomics such as SFS, dNdS and McDonald-Kreitman test to reveal the action of positive selection in B cell clonal lineages. Obviously we observe only the most abundant fraction of real B cell repertoire in a limited period of time in a small cohort group. Nevertheless, evolutionary analysis showed itself as a prominent approach for understanding the biology of B cell immunity. We know just a few of studies that considered B cell repertoires from the phylogenetic point of view, and all of them are focused on immune response to vaccines or viral infections (Nourmohammad et al. 2019; Horns et al. 2019; Hoehn et al. 2021).

In the second part of the work we considered interaction between host and adaptive immunity from the point of view of the pathogen. We tracked the evolution of SARS-CoV-2 inside the immunocompromised host with non-Hodgkin's lymphoma, which lasted almost a year and resulted in accumulation of 42 changes in the viral genome. Surprisingly, it was forced by viral escape from adaptive immunity, in this time presented by cytotoxic T cells ( Chapter 4). As a result, this case study revealed that CD8 T cell escape may be an underappreciated driving force of SARS-CoV-2 evolution. The role of T cell escape in global SARS-CoV-2 evolution still remains an open question.

Thus rapid intrahost evolution can be observed from both sides of the coevolving host-pathogens system. Both hyper mutating B cell clonal lineages and population of the pathogen diversify and adapt right during the time of host-pathogen interaction. Summing up these two studies we have come to the following conclusions:

- Evolutionary regimes of both maturating B cell clonal lineages and population of the pathogen, presented by SAR-CoV-2 in this work, strongly dependent from the second side of host-pathogens system;
- B cell immune response may use different evolutionary regimes depending on the dynamics of interaction with the antigen: persisting B cell memory, showing weak interaction with the antigen by production of BCRs with mostly non-switched IgM/D isotypes, and evolving under relaxed negative selection. In contrast rapidly-expanding antibody producing lineages, probably undergone recent antigen-challenge, show signs of excess of both positive and negative selection in comparison with persisting memory;
- Some of antibody-producing lineages may originate from B cell immune memory reactivation, accompanied with new rounds of coevolution to corresponding antigen by involvement in new cycles of affinity maturation;
- Special conditions of host immune system fully determined the way of SARS-CoV-2 evolution: immune depleted therapy left the patient lack of B cell immunity, thus immune escape from cytotoxic T cells became the major driver of intrahost evolution of SARS-CoV-2;
- The effect of SARS-CoV-2 immune escape was strongly pronounced for a particular host and was not universal for the general population, showing again that intrahost pathogen evolution strongly depends on the second side of host-pathogen interaction.

This work reveals deep interdependency between the work of adaptive immunity and intrahost pathogen reproduction. It proves that these processes should be studied taking into account the fact that they belong to a single interconnected and coevolving system. Evolutionary and phylogenetic analysis of its elements, such as B cell clonal lineages or intrahost viral populations, showed itself as a powerful tool for understanding host-pathogen arms race.

# BIBLIOGRAPHY

Agarwal, Vineet, A. J. Venkatakrishnan, Arjun Puranik, Christian Kirkup, Agustin Lopez-Marquez, Douglas W. Challener, Elitza S. Theel, et al. 2020. "Long-Term SARS-CoV-2 RNA Shedding and Its Temporal Association to IgG Seropositivity." *Cell Death Discovery* 6 (1): 138. https://doi.org/10.1038/s41420-020-00375-y.

Agerer, Benedikt, Maximilian Koblischke, Venugopal Gudipati, Luis Fernando Montaño-Gutierrez, Mark Smyth, Alexandra Popa, Jakob-Wendelin Genger, et al. 2021. "SARS-CoV-2 Mutations in MHC-I-Restricted Epitopes Evade CD8 $^{+}$ T Cell Responses." *Science Immunology* 6 (57): eabg6461. https://doi.org/10.1126/sciimmunol.abg6461.

Allen, Todd M., Marcus Altfeld, Xu G. Yu, Kristin M. O'Sullivan, Mathias Lichterfeld, Sylvie Le Gall, Mina John, et al. 2004. "Selection, Transmission, and Reversion of an Antigen-Processing Cytotoxic T-Lymphocyte Escape Mutation in Human Immunodeficiency Virus Type 1 Infection." *Journal of Virology* 78 (13): 7069–78. https://doi.org/10.1128/JVI.78.13.7069-7078.2004.

Ambrosioni, Juan, José Luis Blanco, Juliana M Reyes-Urueña, Mary-Ann Davies, Omar Sued, Maria Angeles Marcos, Esteban Martínez, et al. 2021. "Overview of SARS-CoV-2 Infection in Adults Living with HIV." *The Lancet HIV* 8 (5): e294–305. https://doi.org/10.1016/S2352-3018(21)00070-9.

Andrew Rambaut, Nick Loman, Oliver Pybus, Wendy Barclay, Jeff Barrett, Alesandro Carabelli, Tom Connor, Tom Peacock, David L Robertson, Erik Volz, on behalf of COVID-19 Genomics Consortium UK (CoG-UK). 2020. "Preliminary Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in the UK Defined by a Novel Set of Spike Mutations.," December. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563/1.

Astuti, Indwiani and Ysrafil. 2020. "Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An Overview of Viral Structure and Host Response." *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14 (4): 407–12. https://doi.org/10.1016/j.dsx.2020.04.020.

Avanzato, Victoria A., M. Jeremiah Matson, Stephanie N. Seifert, Rhys Pryce, Brandi N. Williamson, Sarah L. Anzick, Kent Barbian, et al. 2020. "Case Study: Prolonged

Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer.” *Cell* 183 (7): 1901-1912.e9. https://doi.org/10.1016/j.cell.2020.10.049.

Barak, Michal, Neta S. Zuckerman, Hanna Edelman, Ron Unger, and Ramit Mehr. 2008. “IgTree©: Creating Immunoglobulin Variable Region Gene Lineage Trees.” *Journal of Immunological Methods* 338 (1–2): 67–74. https://doi.org/10.1016/j.jim.2008.06.006.

Bashford-Rogers, Rachael J.M., Anne L. Palser, Brian J. Huntly, Richard Rance, George S. Vassiliou, George A. Follows, and Paul Kellam. 2013. “Network Properties Derived from Deep Sequencing of Human B-Cell Receptor Repertoires Delineate B-Cell Populations.” *Genome Research* 23 (11): 1874–84. https://doi.org/10.1101/gr.154815.113.

Betrains, Albrecht, Laurent Godinas, F. J. Sherida H. Woei‐A‐Jin, Wouter Rosseels, Yannick Van Herck, Natalie Lorent, Daan Dierickx, et al. 2021. “Convalescent Plasma Treatment of Persistent Severe Acute Respiratory Syndrome Coronavirus‐2 (SARS‐CoV‐2) Infection in Patients with Lymphoma with Impaired Humoral Immunity and Lack of Neutralising Antibodies.” *British Journal of Haematology* 192 (6): 1100–1105. https://doi.org/10.1111/bjh.17266.

Boehm, Thomas. 2011. “Design Principles of Adaptive Immune Systems.” *Nature Reviews Immunology* 11 (5): 307–17. https://doi.org/10.1038/nri2944.

Boehm, Thomas, Norimasa Iwanami, and Isabell Hess. 2012. “Evolution of the Immune System in the Lower Vertebrates.” *Annual Review of Genomics and Human Genetics* 13 (1): 127–49. https://doi.org/10.1146/annurev-genom-090711-163747.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Bolotin, Dmitriy A, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, and Dmitriy M Chudakov. 2015. “MiXCR: Software for Comprehensive Adaptive Immunity Profiling.” *Nature Methods* 12 (5): 380–81. https://doi.org/10.1038/nmeth.3364.

Bolthausen, E., and A.-S. Sznitman. 1998. “On Ruelle’s Probability Cascades and an Abstract Cavity Method.” *Communications in Mathematical Physics* 197 (2): 247–76. https://doi.org/10.1007/s002200050450.

Bonsignori, Mattia, Hua-Xin Liao, Feng Gao, Wilton B. Williams, S. Munir Alam, David C. Montefiori, and Barton F. Haynes. 2017. “Antibody-Virus Co-Evolution in HIV Infection: Paths for HIV Vaccine Development.” *Immunological Reviews* 275 (1): 145–60. https://doi.org/10.1111/imr.12509.

Borges, Vítor, Joana Isidro, Mário Cunha, Daniela Cochicho, Luís Martins, Luís Banha, Margarida Figueiredo, et al. 2021. “Long-Term Evolution of SARS-CoV-2 in an Immunocompromised Patient with Non-Hodgkin Lymphoma.” Edited by W. Paul Duprex. *MSphere* 6 (4): e00244-21. https://doi.org/10.1128/mSphere.00244-21.

Briney, Bryan, Anne Inderbitzin, Collin Joyce, and Dennis R. Burton. 2019. “Commonality despite Exceptional Diversity in the Baseline Human Antibody Repertoire.” *Nature* 566 (7744): 393–97. https://doi.org/10.1038/s41586-019-0879-y.

Bronke, Corine, Coral-Ann M. Almeida, Elizabeth McKinnon, Steven G. Roberts, Niamh M. Keane, Abha Chopra, Jonathan M. Carlson, David Heckerman, Simon Mallal, and Mina John. 2013. “HIV Escape Mutations Occur Preferentially at HLA-Binding Sites of CD8 T-Cell Epitopes.” *AIDS* 27 (6): 899–905. https://doi.org/10.1097/QAD.0b013e32835e1616.

Buchmann, Kurt. 2014. “Evolution of Innate Immunity: Clues from Invertebrates via Fish to Mammals.” *Frontiers in Immunology* 5 (September). https://doi.org/10.3389/fimmu.2014.00459.

Buhler, Stéphane, and Alicia Sanchez-Mazas. 2011. “HLA DNA Sequence Variation among Human Populations: Molecular Signatures of Demographic and Selective Events.” Edited by Igor Mokrousov. *PLoS ONE* 6 (2): e14643. https://doi.org/10.1371/journal.pone.0014643.

“Cancer Therapy Adviser.” n.d. https://www.cancertherapyadvisor.com/home/cancer-topics/hematologic-cancers/hematologic-cancers-treatment-regimens/non-hodgkin-lymphoma-nhl-treatment-regimens-diffuse-large-b-cell-lymphoma/.

Cele, Sandile, Farina Karim, Gila Lustig, James Emmanuel San, Tandile Hermanus, Houriiyah Tegally, Jumari Snyman, et al. 2021. “SARS-CoV-2 Evolved during Advanced HIV Disease Immunosuppression Has Beta-like Escape of Vaccine and Delta Infection Elicited Immunity.” Preprint. Infectious Diseases (except HIV/AIDS).

https://doi.org/10.1101/2021.09.14.21263564.

Chakraborty, Chiranjib, Ashish Ranjan Sharma, Manojit Bhattacharya, and Sang-Soo Lee. 2022. “A Detailed Overview of Immune Escape, Antibody Escape, Partial Vaccine Escape of SARS-CoV-2 and Their Emerging Variants With Escape Mutations.” *Frontiers in Immunology* 13 (February): 801522. https://doi.org/10.3389/fimmu.2022.801522.

Chen, Jiahui, Rui Wang, Menglun Wang, and Guo-Wei Wei. 2020. “Mutations Strengthened SARS-CoV-2 Infectivity.” *Journal of Molecular Biology* 432 (19): 5212–26. https://doi.org/10.1016/j.jmb.2020.07.009.

Chen, Jiahui, Rui Wang, and Guo-Wei Wei. 2021. “Review of the Mechanisms of SARS-CoV-2 Evolution and Transmission.” arXiv. http://arxiv.org/abs/2109.08148.

Chen, Xinhua, Zhiyuan Chen, Andrew S. Azman, Ruijia Sun, Wanying Lu, Nan Zheng, Jiaxin Zhou, et al. 2021. “Comprehensive Mapping of Neutralizing Antibodies against SARS-CoV-2 Variants Induced by Natural Infection or Vaccination.” *MedRxiv: The Preprint Server for Health Sciences*, May, 2021.05.03.21256506. https://doi.org/10.1101/2021.05.03.21256506.

Chi, Xiying, Yue Li, and Xiaoyan Qiu. 2020. “V(D)J Recombination, Somatic Hypermutation and Class Switch Recombination of Immunoglobulins: Mechanism and Regulation.” *Immunology* 160 (3): 233–47. https://doi.org/10.1111/imm.13176.

Choi, Bina, Manish C. Choudhary, James Regan, Jeffrey A. Sparks, Robert F. Padera, Xueting Qiu, Isaac H. Solomon, et al. 2020. “Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host.” *New England Journal of Medicine* 383 (23): 2291–93. https://doi.org/10.1056/NEJMc2031364.

Ciurea, Adrian, Lukas Hunziker, Rolf M. Zinkernagel, and Hans Hengartner. 2001. “Viral Escape from the Neutralizing Antibody Response: The Lymphocytic Choriomeningitis Virus Model.” *Immunogenetics* 53 (3): 185–89. https://doi.org/10.1007/s002510100314.

Cohen, Kristen W., Susanne L. Linderman, Zoe Moodie, Julie Czartoski, Lilin Lai, Grace Mantus, Carson Norwood, et al. 2021. “Longitudinal Analysis Shows Durable and Broad Immune Memory after SARS-CoV-2 Infection with Persisting Antibody Responses and Memory B and T Cells.” *Cell Reports Medicine* 2 (7): 100354. https://doi.org/10.1016/j.xcrm.2021.100354.

Connolly, Sarah A., Theodore S. Jardetzky, and Richard Longnecker. 2021. “The Structural Basis of Herpesvirus Entry." *Nature Reviews Microbiology* 19 (2): 110–21. https://doi.org/10.1038/s41579-020-00448-w.

Cooper, Max D., and Matthew N. Alder. 2006. "The Evolution of Adaptive Immune Systems." *Cell* 124 (4): 815–22. https://doi.org/10.1016/j.cell.2006.02.001.

Corcoran, Martin M., Ganesh E. Phad, Néstor Vázquez Bernat, Christiane Stahl-Hennig, Noriyuki Sumida, Mats A.A. Persson, Marcel Martin, and Gunilla B. Karlsson Hedestam. 2016. "Production of Individualized V Gene Databases Reveals High Levels of Immunoglobulin Genetic Diversity." *Nature Communications* 7 (1): 13642. https://doi.org/10.1038/ncomms13642.

Crotty, Shane. 2011. "Follicular Helper CD4 T Cells (T $_{FH}$ )." *Annual Review of Immunology* 29 (1): 621–63. https://doi.org/10.1146/annurev-immunol-031210-101400.

Cunha, Marielton dos Passos, Ana Paula Pessoa Vilela, Camila Vieira Molina, Stephanie Maia Acuña, Sandra Marcia Muxel, Vinícius de Morais Barroso, Sabrina Baroni, et al. 2021. "Atypical Prolonged Viral Shedding With Intra-Host SARS-CoV-2 Evolution in a Mildly Affected Symptomatic Patient." *Frontiers in Medicine* 8 (November): 760170. https://doi.org/10.3389/fmed.2021.760170.

Davidsen, Kristian, and Frederick A. Matsen. 2018. "Benchmarking Tree and Ancestral Sequence Inference for B Cell Receptor Sequences." *Frontiers in Immunology* 9 (October): 2451. https://doi.org/10.3389/fimmu.2018.02451.

Deng, Xianding, Miguel A Garcia-Knight, Mir M. Khalid, Venice Servellita, Candace Wang, Mary Kate Morris, Alicia Sotomayor-González, et al. 2021. "Transmission, Infectivity, and Antibody Neutralization of an Emerging SARS-CoV-2 Variant in California Carrying a L452R Spike Protein Mutation." Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2021.03.07.21252647.

Desjardins, Michel, Mathieu Houde, and Etienne Gagnon. 2005. "Phagocytosis: The Convoluted Way from Nutrition to Adaptive Immunity." *Immunological Reviews* 207 (1): 158–65. https://doi.org/10.1111/j.0105-2896.2005.00319.x.

Dingens, Adam S., Dana Arenz, Haidyn Weight, Julie Overbaugh, and Jesse D. Bloom. 2019. "An Antigenic Atlas of HIV-1 Escape from Broadly Neutralizing Antibodies Distinguishes Functional and Structural Epitopes." *Immunity* 50 (2): 520-532.e3. https://doi.org/10.1016/j.immuni.2018.12.017.

Dolton, Garry, Cristina Rius, Md Samiul Hasan, Barbara Szomolay, Enas Behiry, Thomas Whalley, Joel Southgate, et al. 2021. “Emergence of Immune Escape at Dominant SARS-CoV-2 Killer T-Cell Epitope.” Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2021.06.21.21259010.

Draenert, Rika, Sylvie Le Gall, Katja J. Pfafferott, Alasdair J. Leslie, Polan Chetty, Christian Brander, Edward C. Holmes, et al. 2004. “Immune Selection for Altered Antigen Processing Leads to Cytotoxic T Lymphocyte Escape in Chronic HIV-1 Infection.” *Journal of Experimental Medicine* 199 (7): 905–15. https://doi.org/10.1084/jem.20031982.

Drake, John W., and John J. Holland. 1999. “Mutation Rates among RNA Viruses.” *Proceedings of the National Academy of Sciences* 96 (24): 13910–13. https://doi.org/10.1073/pnas.96.24.13910.

Drummond, Alexei J, Simon Y. W Ho, Matthew J Phillips, and Andrew Rambaut. 2006. “Relaxed Phylogenetics and Dating with Confidence.” Edited by David Penny. *PLoS Biology* 4 (5): e88. https://doi.org/10.1371/journal.pbio.0040088.

Duffy, Siobain. 2018. “Why Are RNA Virus Mutation Rates so Damn High?” *PLOS Biology* 16 (8): e3000003. https://doi.org/10.1371/journal.pbio.3000003.

Dunin-Horkawicz, Stanislaw, Klaus O. Kopec, and Andrei N. Lupas. 2014. “Prokaryotic Ancestry of Eukaryotic Protein Networks Mediating Innate Immunity and Apoptosis.” *Journal of Molecular Biology* 426 (7): 1568–82. https://doi.org/10.1016/j.jmb.2013.11.030.

Dutta, Abhishek. 2022. “COVID-19 Waves: Variant Dynamics and Control.” *Scientific Reports* 12 (1): 9332. https://doi.org/10.1038/s41598-022-13371-2.

Edgar, R. C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. https://doi.org/10.1093/nar/gkh340.

Emery, Madison A., Bradford A. Dimos, and Laura D. Mydlarz. 2021. “Cnidarian Pattern Recognition Receptor Repertoires Reflect Both Phylogeny and Life History Traits.” *Frontiers in Immunology* 12 (June): 689463. https://doi.org/10.3389/fimmu.2021.689463.

Erdmann, Nathan, Victor Y. Du, Jonathan Carlson, Malinda Schaefer, Alexander Jureka, Sarah Sterrett, Ling Yue, et al. 2015. “HLA Class-II Associated HIV Polymorphisms Predict

Escape from CD4+ T Cell Responses." Edited by Ronald Swanstrom. *PLOS Pathogens* 11 (8): e1005111. https://doi.org/10.1371/journal.ppat.1005111.

Erickson, Ann L, Yoichi Kimura, Suzu Igarashi, Jennifer Eichelberger, Michael Houghton, John Sidney, Denise McKinney, Alessandro Sette, Austin L Hughes, and Christopher M Walker. 2001. "The Outcome of Hepatitis C Virus Infection Is Predicted by Escape Mutations in Epitopes Targeted by Cytotoxic T Lymphocytes." *Immunity* 15 (6): 883–95. https://doi.org/10.1016/S1074-7613(01)00245-X.

Eschli, Bruno, Raphaël M. Zellweger, Alexander Wepf, Karl S. Lang, Katharina Quirin, Jacqueline Weber, Rolf M. Zinkernagel, and Hans Hengartner. 2007. "Early Antibodies Specific for the Neutralizing Epitope on the Receptor Binding Subunit of the Lymphocytic Choriomeningitis Virus Glycoprotein Fail To Neutralize the Virus." *Journal of Virology* 81 (21): 11650–57. https://doi.org/10.1128/JVI.00955-07.

Feng, Dan, Sake J. de Vlas, Li-Qun Fang, Xiao-Na Han, Wen-Juan Zhao, Shen Sheng, Hong Yang, Zhong-Wei Jia, Jan Hendrik Richardus, and Wu-Chun Cao. 2009. "The SARS Epidemic in Mainland China: Bringing Together All Epidemiological Data." *Tropical Medicine & International Health* 14 (November): 4–13. https://doi.org/10.1111/j.1365-3156.2008.02145.x.

Ferretti, Andrew P., Tomasz Kula, Yifan Wang, Dalena M.V. Nguyen, Adam Weinheimer, Garrett S. Dunlap, Qikai Xu, et al. 2020. "Unbiased Screens Show CD8+ T Cells of COVID-19 Patients Recognize Shared Epitopes in SARS-CoV-2 That Largely Reside Outside the Spike Protein." *Immunity* 53 (5): 1095-1107.e3. https://doi.org/10.1016/j.immuni.2020.10.006.

Flajnik, Martin F., and Masanori Kasahara. 2010. "Origin and Evolution of the Adaptive Immune System: Genetic Events and Selective Pressures." *Nature Reviews Genetics* 11 (1): 47–59. https://doi.org/10.1038/nrg2703.

Focosi, Daniele, and Fabrizio Maggi. 2021. "Neutralising Antibody Escape of SARS‐CoV‐2 Spike Protein: Risk Assessment for Antibody‐based Covid‐19 Therapeutics and Vaccines." *Reviews in Medical Virology*, March, rmv.2231. https://doi.org/10.1002/rmv.2231.

Fu, Yu, Ping Han, Rui Zhu, Tao Bai, Jianhua Yi, Xi Zhao, Meihui Tao, et al. 2020. "Risk Factors for Viral RNA Shedding in COVID-19 Patients." *European Respiratory Journal* 56 (1):

2001190. https://doi.org/10.1183/13993003.01190-2020.

Fujita, Teizo. 2002. “Evolution of the Lectin–Complement Pathway and Its Role in Innate Immunity.” *Nature Reviews Immunology* 2 (5): 346–53. https://doi.org/10.1038/nri800.

Gadala-Maria, Daniel, Gur Yaari, Mohamed Uduman, and Steven H. Kleinstein. 2015. “Automated Analysis of High-Throughput B-Cell Sequencing Data Reveals a High Frequency of Novel Immunoglobulin V Gene Segment Alleles.” *Proceedings of the National Academy of Sciences* 112 (8): E862–70. https://doi.org/10.1073/pnas.1417683112.

Gangaev, Anastasia, Steven L. C. Ketelaars, Olga I. Isaeva, Sanne Patiwael, Anna Dopler, Kelly Hoefakker, Sara De Biasi, et al. 2021. “Identification and Characterization of a SARS-CoV-2 Specific CD8+ T Cell Response with Immunodominant Features.” *Nature Communications* 12 (1): 2593. https://doi.org/10.1038/s41467-021-22811-y.

Gangavarapu, Karthik, Alaa Abdel Latif, Julia L. Mullen, Manar Alkuzweny, Emory Hufbauer, Ginger Tsueng, Emily Haag, et al. 2022. “Outbreak.Info Genomic Reports: Scalable and Dynamic Surveillance of SARS-CoV-2 Variants and Mutations.” Preprint. Epidemiology. https://doi.org/10.1101/2022.01.27.22269965.

Gao, Fan, and Kai Wang. 2015. “Ligation-Anchored PCR Unveils Immune Repertoire of TCR-Beta from Whole Blood.” *BMC Biotechnology* 15 (1): 39. https://doi.org/10.1186/s12896-015-0153-9.

Garcia, K. Christopher, and Erin J. Adams. 2005. “How the T Cell Receptor Sees Antigen—A Structural View.” *Cell* 122 (3): 333–36. https://doi.org/10.1016/j.cell.2005.07.015.

Garrett, Meghan E., Jared Galloway, Helen Y. Chu, Hannah L. Itell, Caitlin I. Stoddard, Caitlin R. Wolf, Jennifer K. Logue, et al. 2021. “High Resolution Profiling of Pathways of Escape for SARS-CoV-2 Spike-Binding Antibodies.” *Cell*, May, S0092867421005808. https://doi.org/10.1016/j.cell.2021.04.045.

Geffard, Estelle, Sophie Limou, Alexandre Walencik, Michelle Daya, Harold Watson, Dara Torgerson, Kathleen C Barnes, et al. 2020. “Easy-HLA: A Validated Web Application Suite to Reveal the Full Details of HLA Typing.” Edited by Alfonso Valencia. *Bioinformatics* 36 (7): 2157–64. https://doi.org/10.1093/bioinformatics/btz875.

Gentles, Lauren E., Hongquan Wan, Maryna C. Eichelberger, and Jesse D. Bloom. 2020. “Antibody Neutralization of an Influenza Virus That Uses Neuraminidase for Receptor

Binding." *Viruses* 12 (6): 597. https://doi.org/10.3390/v12060597.

Geretti, Anna Maria, and Tomas Doyle. 2010. "Immunization for HIV-Positive Individuals:" *Current Opinion in Infectious Diseases* 23 (1): 32–38. https://doi.org/10.1097/QCO.0b013e328334fec4.

Ginestet, Cedric. 2011. "Ggplot2: Elegant Graphics for Data Analysis: Book Reviews." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174 (1): 245–46. https://doi.org/10.1111/j.1467-985X.2010.00676_9.x.

Gojobori, Nei. 1986. "Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions." *Molecular Biology and Evolution*, September. https://doi.org/10.1093/oxfordjournals.molbev.a040410.

Gong, Shang Yu, Debashree Chatterjee, Jonathan Richard, Jérémie Prévost, Alexandra Tauzin, Romain Gasser, Yuxia Bo, et al. 2021. "Contribution of Single Mutations to Selected SARS-CoV-2 Emerging Variants Spike Antigenicity." *Virology* 563 (November): 134–45. https://doi.org/10.1016/j.virol.2021.09.001.

Goulder, Philip J. R., Christian Brander, Yanhua Tang, Cecile Tremblay, Robert A. Colbert, Marylyn M. Addo, Eric S. Rosenberg, et al. 2001. "Evolution and Transmission of Stable CTL Escape Mutations in HIV Infection." *Nature* 412 (6844): 334–38. https://doi.org/10.1038/35085576.

Graudenzi, Alex, Davide Maspero, Fabrizio Angaroni, Rocco Piazza, and Daniele Ramazzotti. 2021. "Mutational Signatures and Heterogeneous Host Response Revealed via Large-Scale Characterization of SARS-CoV-2 Genomic Diversity." *IScience* 24 (2): 102116. https://doi.org/10.1016/j.isci.2021.102116.

Gribble, Jennifer, Laura J. Stevens, Maria L. Agostini, Jordan Anderson-Daniels, James D. Chappell, Xiaotao Lu, Andrea J. Pruijssers, Andrew L. Routh, and Mark R. Denison. 2021. "The Coronavirus Proofreading Exoribonuclease Mediates Extensive Viral Recombination." Edited by Andrew Pekosz. *PLOS Pathogens* 17 (1): e1009226. https://doi.org/10.1371/journal.ppat.1009226.

Grubaugh, Nathan D., Karthik Gangavarapu, Joshua Quick, Nathaniel L. Matteson, Jaqueline Goes De Jesus, Bradley J. Main, Amanda L. Tan, et al. 2019. "An Amplicon-Based Sequencing Framework for Accurately Measuring Intrahost Virus Diversity Using PrimalSeq and IVar." *Genome Biology* 20 (1): 8.

https://doi.org/10.1186/s13059-018-1618-7.

Guo, Wei, Fangzhao Ming, Yu Dong, Qian Zhang, Xiaoxia Zhang, Pingzheng Mo, Yong Feng, and Ke Liang. 2020. “A Survey for COVID-19 Among HIV/AIDS Patients in Two Districts of Wuhan, China.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3550029.

Gupta, Namita T., Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H. Kleinstein. 2015. “Change-O: A Toolkit for Analyzing Large-Scale B Cell Immunoglobulin Repertoire Sequencing Data: Table 1.” *Bioinformatics* 31 (20): 3356–58. https://doi.org/10.1093/bioinformatics/btv359.

Habel, Jennifer R, Thi H O Nguyen, Carolien E van de Sandt, Jennifer A Juno, Priyanka Chaurasia, Kathleen Wragg, Marios Koutsakos, et al. 2020. “Suboptimal SARS-CoV-2-Specific CD8 $^{+}$ T-Cell Response Associated with the Prominent HLA-A*02:01 Phenotype.” Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.08.17.20176370.

Hahn, Thomas von, Joo Chun Yoon, Harvey Alter, Charles M. Rice, Barbara Rehermann, Peter Balfe, and Jane A. McKeating. 2007. “Hepatitis C Virus Continuously Escapes From Neutralizing Antibody and T-Cell Responses During Chronic Infection In Vivo.” *Gastroenterology* 132 (2): 667–78. https://doi.org/10.1053/j.gastro.2006.12.008.

Halaji, Mehrdad, Mohammad Heiat, Niloofar Faraji, and Reza Ranjbar. 2021. “Epidemiology of COVID-19: An Updated Review.” *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences* 26: 82. https://doi.org/10.4103/jrms.JRMS_506_20.

Hall, Matthew D., Mark E. J. Woolhouse, and Andrew Rambaut. 2016. “The Effects of Sampling Strategy on the Quality of Reconstruction of Viral Population Dynamics Using Bayesian Skyline Family Coalescent Methods: A Simulation Study.” *Virus Evolution* 2 (1). https://doi.org/10.1093/ve/vew003.

Harvey, William T., Alessandro M. Carabelli, Ben Jackson, Ravindra K. Gupta, Emma C. Thomson, Ewan M. Harrison, Catherine Ludden, et al. 2021. “SARS-CoV-2 Variants, Spike Mutations and Immune Escape.” *Nature Reviews. Microbiology* 19 (7): 409–24. https://doi.org/10.1038/s41579-021-00573-0.

Heather, James M., Katharine Best, Theres Oakes, Eleanor R. Gray, Jennifer K. Roe, Niclas

Thomas, Nir Friedman, Mahdad Noursadeghi, and Benjamin Chain. 2016. “Dynamic Perturbations of the T-Cell Receptor Repertoire in Chronic HIV Infection and Following Antiretroviral Therapy.” *Frontiers in Immunology* 6 (January). https://doi.org/10.3389/fimmu.2015.00644.

Heesters, Balthasar A., Cees E. van der Poel, Abhishek Das, and Michael C. Carroll. 2016. “Antigen Presentation to B Cells.” *Trends in Immunology* 37 (12): 844–54. https://doi.org/10.1016/j.it.2016.10.003.

Hodcroft Emma. 2020. “CoVariants.” 2020. https://covariants.org/.

Hoehn, Kenneth B., Jackson S. Turner, Frederick I. Miller, Ruoyi Jiang, Oliver G. Pybus, Ali H. Ellebedy, and Steven H. Kleinstein. 2021. “Human B Cell Lineages Engaged by Germinal Centers Following Influenza Vaccination Are Measurably Evolving.” Preprint. Immunology. https://doi.org/10.1101/2021.01.06.425648.

Horns, Felix, Christopher Vollmers, Derek Croote, Sally F Mackey, Gary E Swan, Cornelia L Dekker, Mark M Davis, and Stephen R Quake. 2016. “Lineage Tracing of Human B Cells Reveals the in Vivo Landscape of Human Antibody Class Switching.” *ELife* 5 (August): e16578. https://doi.org/10.7554/eLife.16578.

Horns, Felix, Christopher Vollmers, Cornelia L. Dekker, and Stephen R. Quake. 2019. “Signatures of Selection in the Human Antibody Repertoire: Selective Sweeps, Competing Subclones, and Neutral Drift.” *Proceedings of the National Academy of Sciences* 116 (4): 1261–66. https://doi.org/10.1073/pnas.1814213116.

Hu, Jie, Pai Peng, Xiaoxia Cao, Kang Wu, Juan Chen, Kai Wang, Ni Tang, and Ai-long Huang. 2022. “Increased Immune Escape of the New SARS-CoV-2 Variant of Concern Omicron.” *Cellular & Molecular Immunology* 19 (2): 293–95. https://doi.org/10.1038/s41423-021-00836-z.

Hueso, Thomas, Cécile Pouderoux, Hélène Péré, Anne-Lise Beaumont, Laure-Anne Raillon, Florence Ader, Lucienne Chatenoud, et al. 2020. “Convalescent Plasma Therapy for B-Cell–Depleted Patients with Protracted COVID-19.” *Blood* 136 (20): 2290–95. https://doi.org/10.1182/blood.2020008423.

Imkeller, Katharina, and Hedda Wardemann. 2018. “Assessing Human B Cell Repertoire Diversity and Convergence.” *Immunological Reviews* 284 (1): 51–66. https://doi.org/10.1111/imr.12670.

Islam, SM Rashed Ul, Tahmina Akther, Sharmin Sultana, and Saif Ullah Munshi. 2021. “Persistence of SARS-CoV-2 RNA in a Male with Metabolic Syndrome for 72 Days: A Case Report.” *SAGE Open Medical Case Reports* 9 (January): 2050313X2198949. https://doi.org/10.1177/2050313X21989492.

Itokawa, Kentaro, Tsuyoshi Sekizuka, Masanori Hashino, Rina Tanaka, and Makoto Kuroda. 2020. “Disentangling Primer Interactions Improves SARS-CoV-2 Genome Sequencing by Multiplex Tiling PCR.” Edited by Ruslan Kalendar. *PLOS ONE* 15 (9): e0239403. https://doi.org/10.1371/journal.pone.0239403.

Jung, David, and Frederick W Alt. 2004. “Unraveling V(D)J Recombination.” *Cell* 116 (2): 299–311. https://doi.org/10.1016/S0092-8674(04)00039-X.

Kapila, Ketoki. 2004. “Kuby Immunology, 4th Edition Year 2000.” *Medical Journal Armed Forces India* 60 (1): 91. https://doi.org/10.1016/S0377-1237(04)80176-X.

Karataş, Ayşe, Ahmet Çağkan İnkaya, Haluk Demiroğlu, Salih Aksu, Tahmaz Haziyev, Olgu Erkin Çınar, Alpaslan Alp, Ömrüm Uzun, Nilgün Sayınalp, and Hakan Göker. 2020. “Prolonged Viral Shedding in a Lymphoma Patient with COVID-19 Infection Receiving Convalescent Plasma.” *Transfusion and Apheresis Science* 59 (5): 102871. https://doi.org/10.1016/j.transci.2020.102871.

Karim, F, Mys Moosa, Bi Gosnell, S Cele, J Giandhari, S Pillay, H Tegally, et al. 2021. “Persistent SARS-CoV-2 Infection and Intra-Host Evolution in Association with Advanced HIV Infection.” Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2021.06.03.21258228.

Katoh, K., and D. M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Kawaguchi, Shuji, and Fumihiko Matsuda. 2020. “High-Definition Genomic Analysis of HLA Genes Via Comprehensive HLA Allele Genotyping.” In *Immunoinformatics*, edited by Namrata Tomar, 2131:31–38. Methods in Molecular Biology. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-0389-5_3.

Kemp, Sa, Da Collier, R Datir, Iatm Ferreira, S Gayed, A Jahun, M Hosmillo, et al. 2020. “Neutralising Antibodies in Spike Mediated SARS-CoV-2 Adaptation.” Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.12.05.20241927.

Khatamzas, Elham, Maximilian Muenchhoff, Alexandra Rehn, Alexander Graf, Johannes Hellmuth, Alexandra Hollaus, Anne-Wiebe Mohr, et al. 2021. “CD8 T Cells and Antibodies Drive SARS-CoV-2 Evolution in Chronic Infection.” Preprint. In Review. https://doi.org/10.21203/rs.3.rs-846197/v1.

Khatamzas, Elham, Alexandra Rehn, Maximilian Muenchhoff, Johannes Hellmuth, Erik Gaitzsch, Tobias Weiglein, Enrico Georgi, et al. 2021. “Emergence of Multiple SARS-CoV-2 Mutations in an Immunocompromised Host.” Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2021.01.10.20248871.

Kim, Sinae, Tam T. Nguyen, Afeisha S. Taitt, Hyunjhung Jhun, Ho-Young Park, Sung-Han Kim, Yong-Gil Kim, et al. 2021. “SARS-CoV-2 Omicron Mutation Is Faster than the Chase: Multiple Mutations on Spike/ACE2 Interaction Residues.” *Immune Network* 21 (6): e38. https://doi.org/10.4110/in.2021.21.e38.

Kim, Yoo-Ah, Mark D.M. Leiserson, Priya Moorjani, Roded Sharan, Damian Wojtowicz, and Teresa M. Przytycka. 2021. “Mutational Signatures: From Methods to Mechanisms.” *Annual Review of Biomedical Data Science* 4 (1): 189–206. https://doi.org/10.1146/annurev-biodatasci-122320-120920.

Kimura, Yoichi, Toshifumi Gushima, Sharad Rawale, Pravin Kaumaya, and Christopher M. Walker. 2005. “Escape Mutations Alter Proteasome Processing of Major Histocompatibility Complex Class I-Restricted Epitopes in Persistent Hepatitis C Virus Infection.” *Journal of Virology* 79 (8): 4870–76. https://doi.org/10.1128/JVI.79.8.4870-4876.2005.

Kingman, J.F.C. 1982. “The Coalescent.” *Stochastic Processes and Their Applications* 13 (3): 235–48. https://doi.org/10.1016/0304-4149(82)90011-4.

Klarenbeek, Paul L., Paul P. Tak, Barbera D.C. van Schaik, Aeilko H. Zwinderman, Marja E. Jakobs, Zhuoli Zhang, Antoine H.C. van Kampen, René A.W. van Lier, Frank Baas, and Niek de Vries. 2010. “Human T-Cell Memory Consists Mainly of Unexpanded Clones.” *Immunology Letters* 133 (1): 42–48. https://doi.org/10.1016/j.imlet.2010.06.011.

Klink, Galya V., Ksenia Safina, Elena Nabieva, Nikita Shvyrev, Sofya Garushyants, Evgeniia Alekseeva, Andrey B. Komissarov, et al. 2021. “The Rise and Spread of the SARS-CoV-2 AY.122 Lineage in Russia.” Preprint. Epidemiology. https://doi.org/10.1101/2021.12.02.21267168.

Knorre, Dmitry, Elena Nabieva, Sofya Garushyants, and The CoRGI (Coronavirus Russian Genetic Initiative) Consortium. 2021. "Taxameter.Ru. Available Online: Http://Taxameter.Ru/ (2021)." 2021. https://taxameter.ru/.

Koh, Gene, Andrea Degasperi, Xueqing Zou, Sophie Momen, and Serena Nik-Zainal. 2021. "Mutational Signatures: Emerging Concepts, Caveats and Clinical Applications." *Nature Reviews Cancer* 21 (10): 619–37. https://doi.org/10.1038/s41568-021-00377-7.

Komissarov, Andrey B., Ksenia R. Safina, Sofya K. Garushyants, Artem V. Fadeev, Mariia V. Sergeeva, Anna A. Ivanova, Daria M. Danilenko, et al. 2021. "Genomic Epidemiology of the Early Stages of the SARS-CoV-2 Outbreak in Russia." *Nature Communications* 12 (1): 649. https://doi.org/10.1038/s41467-020-20880-z.

Korber, Bette, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, et al. 2020. "Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus." *Cell* 182 (4): 812-827.e19. https://doi.org/10.1016/j.cell.2020.06.043.

Kosakovsky Pond, Sergei L., and Simon D. W. Frost. 2005. "Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection." *Molecular Biology and Evolution* 22 (5): 1208–22. https://doi.org/10.1093/molbev/msi105.

Koyama, Takahiko, Daniel Platt, and Laxmi Parida. 2020. "Variant Analysis of SARS-CoV-2 Genomes." *Bulletin of the World Health Organization* 98 (7): 495–504. https://doi.org/10.2471/BLT.20.253591.

Krammer, Florian. 2019. "The Human Antibody Response to Influenza A Virus Infection and Vaccination." *Nature Reviews Immunology* 19 (6): 383–97. https://doi.org/10.1038/s41577-019-0143-6.

Kräutler, Nike Julia, Alexander Yermanos, Alessandro Pedrioli, Suzanne P.M. Welten, Dominique Lorgé, Ute Greczmiel, Ilka Bartsch, et al. 2020. "Quantitative and Qualitative Analysis of Humoral Immunity Reveals Continued and Personalized Evolution in Chronic Viral Infection." *Cell Reports* 30 (4): 997-1012.e6. https://doi.org/10.1016/j.celrep.2019.12.088.

Kupferschmidt, Kai. 2021. "Where Did 'Weird' Omicron Come From?" *Science* 374 (6572): 1179–1179. https://doi.org/10.1126/science.acx9738.

Küppers, Ralf. 2005. "Mechanisms of B-Cell Lymphoma Pathogenesis." *Nature Reviews Cancer*

5 (4): 251–62. https://doi.org/10.1038/nrc1589.

Lazarevic, Ivana, Ana Banko, Danijela Miljanovic, and Maja Cupic. 2019. “Immune-Escape Hepatitis B Virus Mutations Associated with Viral Reactivation upon Immunosuppression.” *Viruses* 11 (9): 778. https://doi.org/10.3390/v11090778.

Lee, Christina Y., Monika K Shah, David Hoyos, Alexander Solovyov, Melanie Douglas, Ying Taur, Peter Maslak, et al. 2021. “Prolonged SARS-CoV-2 Infection in Patients with Lymphoid Malignancies.” *Cancer Discovery*, November, candisc.1033.2021. https://doi.org/10.1158/2159-8290.CD-21-1033.

Lee, Robin D., Sarah A. Munro, Todd P. Knutson, Rebecca S. LaRue, Lynn M. Heltemes-Harris, and Michael A. Farrar. 2021. “Single-Cell Analysis Identifies Dynamic Gene Expression Networks That Govern B Cell Development and Transformation.” *Nature Communications* 12 (1): 6843. https://doi.org/10.1038/s41467-021-27232-5.

Leon, Paul E., Teddy John Wohlbold, Wenqian He, Mark J. Bailey, Carole J. Henry, Patrick C. Wilson, Florian Krammer, and Gene S. Tan. 2017. “Generation of Escape Variants of Neutralizing Influenza Virus Monoclonal Antibodies.” *Journal of Visualized Experiments*, no. 126 (August): 56067. https://doi.org/10.3791/56067.

Letunic, Ivica, and Peer Bork. 2007. “Interactive Tree Of Life (ITOL): An Online Tool for Phylogenetic Tree Display and Annotation.” *Bioinformatics (Oxford, England)* 23 (1): 127–28. https://doi.org/10.1093/bioinformatics/btl529.

Leung, Wayne F., Samuel Chorlton, John Tyson, Ghada N. Al-Rawahi, Agatha N. Jassem, Natalie Prystajecky, Shazia Masud, et al. 2022. “COVID-19 in an Immunocompromised Host: Persistent Shedding of Viable SARS-CoV-2 and Emergence of Multiple Mutations: A Case Report.” *International Journal of Infectious Diseases* 114 (January): 178–82. https://doi.org/10.1016/j.ijid.2021.10.045.

Li, H. 2011. “A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data.” *Bioinformatics* 27 (21): 2987–93. https://doi.org/10.1093/bioinformatics/btr509.

Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *ArXiv:1303.3997 [q-Bio]*, May. http://arxiv.org/abs/1303.3997.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing

Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Liu, Xiao, and Jinghua Wu. 2018. “History, Applications, and Challenges of Immune Repertoire Research.” *Cell Biology and Toxicology* 34 (6): 441–57. https://doi.org/10.1007/s10565-018-9426-0.

Lord, Jennifer S, and Michael B Bonsall. 2021. “The Evolutionary Dynamics of Viruses: Virion Release Strategies, Time Delays and Fitness Minima.” *Virus Evolution* 7 (1): veab039. https://doi.org/10.1093/ve/veab039.

Lucas, Michaela, Urs Karrer, Andrew Lucas, and Paul Klenerman. 2008. “Viral Escape Mechanisms - Escapology Taught by Viruses: Viral Escape Mechanisms.” *International Journal of Experimental Pathology* 82 (5): 269–86. https://doi.org/10.1046/j.1365-2613.2001.00204.x.

Lyngse, Frederik Plesner, Kåre Mølbak, Robert Leo Skov, Lasse Engbo Christiansen, Laust Hvas Mortensen, Mads Albertsen, Camilla Holten Møller, et al. 2021. “Increased Transmissibility of SARS-CoV-2 Lineage B.1.1.7 by Age and Viral Load.” *Nature Communications* 12 (1): 7251. https://doi.org/10.1038/s41467-021-27202-x.

Mamedov, Ilgar Z., Olga V. Britanova, Dmitriy A. Bolotin, Anna V. Chkalina, Dmitriy B. Staroverov, Ivan V. Zvyagin, Alexey A. Kotlobay, et al. 2011. “Quantitative Tracking of T Cell Clones after Haematopoietic Stem Cell Transplantation.” *EMBO Molecular Medicine* 3 (4): 201–7. https://doi.org/10.1002/emmm.201100129.

Mamedov, Ilgar Z., Olga V. Britanova, Ivan V. Zvyagin, Maria A. Turchaninova, Dmitriy A. Bolotin, Ekaterina V. Putintseva, Yuriy B. Lebedev, and Dmitriy M. Chudakov. 2013. “Preparing Unbiased T-Cell Receptor and Antibody CDNA Libraries for the Deep Next Generation Sequencing Profiling.” *Frontiers in Immunology* 4. https://doi.org/10.3389/fimmu.2013.00456.

Marchalonis, John J., Ingvill Jensen, and Samuel F. Schluter. 2002. “Structural, Antigenic and Evolutionary Analyses of Immunoglobulins and T Cell Receptors.” *Journal of Molecular Recognition* 15 (5): 260–71. https://doi.org/10.1002/jmr.586.

Maróstica, André Silva, Kelly Nunes, Erick C. Castelli, Nayane S. B. Silva, Bruce S. Weir, Jérôme Goudet, and Diogo Meyer. 2022. “How HLA Diversity Is Apportioned: Influence of Selection and Relevance to Transplantation.” *Philosophical Transactions of the Royal*

*Society B: Biological Sciences* 377 (1852): 20200420. https://doi.org/10.1098/rstb.2020.0420.

Marty, Rachel, Saghar Kaabinejadian, David Rossell, Michael J. Slifker, Joris van de Haar, Hatice Billur Engin, Nicola de Prisco, et al. 2017. "MHC-I Genotype Restricts the Oncogenic Mutational Landscape." *Cell* 171 (6): 1272-1283.e15. https://doi.org/10.1016/j.cell.2017.09.050.

McCarthy, Kevin R., Donald D. Raymond, Khoi T. Do, Aaron G. Schmidt, and Stephen C. Harrison. 2019. "Affinity Maturation in a Human Humoral Response to Influenza Hemagglutinin." *Proceedings of the National Academy of Sciences* 116 (52): 26745–51. https://doi.org/10.1073/pnas.1915620116.

McCarthy, Kevin R., Linda J. Rennick, Sham Nambulli, Lindsey R. Robinson-McCarthy, William G. Bain, Ghady Haidar, and W. Paul Duprex. 2021. "Recurrent Deletions in the SARS-CoV-2 Spike Glycoprotein Drive Antibody Escape." *Science* 371 (6534): 1139–42. https://doi.org/10.1126/science.abf6950.

McCoy, Connor O., Trevor Bedford, Vladimir N. Minin, Philip Bradley, Harlan Robins, and Frederick A. Matsen. 2015. "Quantifying Evolutionary Constraints on B-Cell Affinity Maturation." *Philosophical Transactions of the Royal Society B: Biological Sciences* 370 (1676): 20140244. https://doi.org/10.1098/rstb.2014.0244.

McDonald, John H., and Martin Kreitman. 1991. "Adaptive Protein Evolution at the Adh Locus in Drosophila." *Nature* 351 (6328): 652–54. https://doi.org/10.1038/351652a0.

Meijers, Matthijs, Kanika Vanshylla, Henning Gruell, Florian Klein, and Michael Lässig. 2021. "Predicting in Vivo Escape Dynamics of HIV-1 from a Broadly Neutralizing Antibody." *Proceedings of the National Academy of Sciences* 118 (30): e2104651118. https://doi.org/10.1073/pnas.2104651118.

Mendiola-Pastrana, Indira R., Eduardo López-Ortiz, José G. Río de la Loza-Zamora, James González, Anel Gómez-García, and Geovani López-Ortiz. 2022. "SARS-CoV-2 Variants and Clinical Outcomes: A Systematic Review." *Life* 12 (2): 170. https://doi.org/10.3390/life12020170.

Mikocziova, Ivana, Victor Greiff, and Ludvig M. Sollid. 2021. "Immunoglobulin Germline Gene Variation and Its Impact on Human Disease." *Genes & Immunity* 22 (4): 205–17. https://doi.org/10.1038/s41435-021-00145-5.

Minervina, Anastasia, Mikhail Pogorelyy, and Ilgar Mamedov. 2019. “T‐cell Receptor and B‐cell Receptor Repertoire Profiling in Adaptive Immunity.” *Transplant International* 32 (11): 1111–23. https://doi.org/10.1111/tri.13475.

Mirzaei, Hossein, Willi McFarland, Mohammad Karamouzian, and Hamid Sharifi. 2021. “COVID-19 Among People Living with HIV: A Systematic Review.” *AIDS and Behavior* 25 (1): 85–92. https://doi.org/10.1007/s10461-020-02983-2.

Mohapatra, Ranjan K., Ruchi Tiwari, Ashish K. Sarangi, Sanjay K. Sharma, Rekha Khandia, G. Saikumar, and Kuldeep Dhama. 2022. “Twin Combination of Omicron and Delta Variants Triggering a Tsunami Wave of Ever High Surges in COVID‐19 Cases: A Challenging Global Threat with a Special Focus on the Indian Subcontinent.” *Journal of Medical Virology* 94 (5): 1761–65. https://doi.org/10.1002/jmv.27585.

Monrad, Ida, Signe Risgaard Sahlertz, Stine Sofie Frank Nielsen, Louise Ørnskov Pedersen, Mikkel Steen Petersen, Carl Mathias Kobel, Irene Harder Tarpgaard, et al. 2021. “Persistent Severe Acute Respiratory Syndrome Coronavirus 2 Infection in Immunocompromised Host Displaying Treatment Induced Viral Evolution.” *Open Forum Infectious Diseases* 8 (7): ofab295. https://doi.org/10.1093/ofid/ofab295.

Moore, Joanna L., Pavan V. Ganapathiraju, Caroline P. Kurtz, and Booth Wainscoat. 2020. “A 63-Year-Old Woman with a History of Non-Hodgkin Lymphoma with Persistent SARS-CoV-2 Infection Who Was Seronegative and Treated with Convalescent Plasma.” *American Journal of Case Reports* 21 (September). https://doi.org/10.12659/AJCR.927812.

Moss, Paul. 2022. “The T Cell Immune Response against SARS-CoV-2.” *Nature Immunology* 23 (2): 186–93. https://doi.org/10.1038/s41590-021-01122-w.

Muecksch, Frauke, Yiska Weisblum, Christopher O. Barnes, Fabian Schmidt, Dennis Schaefer-Babajew, Zijun Wang, Julio C. C. Lorenzi, et al. 2021. “Affinity Maturation of SARS-CoV-2 Neutralizing Antibodies Confers Potency, Breadth, and Resilience to Viral Escape Mutations.” *Immunity* 54 (8): 1853-1868.e7. https://doi.org/10.1016/j.immuni.2021.07.008.

Mukhina, Olgo A., Daria S. Fomina, Vasiliy V. Parshin, Vladimir A. Gushchin, Inna V. Dolzhikova, Alexey M. Shchetinin, Dmitriy M. Chudakov, et al. 2022. “SARS-CoV-2 Evolution in a Patient with Secondary B-Cell Immunodeficiency: A Clinical Case.”

*Human Vaccines & Immunotherapeutics*, August, 2101334. https://doi.org/10.1080/21645515.2022.2101334.

Mulder, Marlies, Dewi S J M van der Vegt, Bas B Oude Munnink, Corine H GeurtsvanKessel, Jeroen van de Bovenkamp, Reina S Sikkema, Esther M G Jacobs, Marion P G Koopmans, and Marjolijn C A Wegdam-Blans. 2021. "Reinfection of Severe Acute Respiratory Syndrome Coronavirus 2 in an Immunocompromised Patient: A Case Report." *Clinical Infectious Diseases* 73 (9): e2841–42. https://doi.org/10.1093/cid/ciaa1538.

Murphy, Kenneth, and Casey Weaver. 2016. *Janeway's Immunobiology*. 9th edition. New York, NY: Garland Science/Taylor & Francis Group, LLC.

Nakajima, Yukiko, Asuca Ogai, Karin Furukawa, Ryosuke Arai, Ryusuke Anan, Yasushi Nakano, Yuko Kurihara, Hideaki Shimizu, Takako Misaki, and Nobuhiko Okabe. 2021. "Prolonged Viral Shedding of SARS-CoV-2 in an Immunocompromised Patient." *Journal of Infection and Chemotherapy* 27 (2): 387–89. https://doi.org/10.1016/j.jiac.2020.12.001.

Neher, Richard A., Taylor A. Kessinger, and Boris I. Shraiman. 2013. "Coalescence and Genetic Diversity in Sexual Populations under Selection." *Proceedings of the National Academy of Sciences* 110 (39): 15836–41. https://doi.org/10.1073/pnas.1309697110.

Neher, Richard A, Colin A Russell, and Boris I Shraiman. 2014. "Predicting Evolution from the Shape of Genealogical Trees." *ELife* 3 (November): e03568. https://doi.org/10.7554/eLife.03568.

Nei, Masatoshi, and Sudhir Kumar. 2000. *Molecular Evolution and Phylogenetics*. Oxford ; New York: Oxford University Press.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.

Nielsen, Rasmus. 2005. "Molecular Signatures of Natural Selection." *Annual Review of Genetics* 39 (1): 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420.

Nourmohammad, Armita, Jakub Otwinowski, Marta Łuksza, Thierry Mora, and Aleksandra M Walczak. 2019. "Fierce Selection and Interference in B-Cell Repertoire Response to

Chronic HIV-1." Edited by Thomas Leitner. *Molecular Biology and Evolution* 36 (10): 2184–94. https://doi.org/10.1093/molbev/msz143.

Odegard, Valerie H., and David G. Schatz. 2006. "Targeting of Somatic Hypermutation." *Nature Reviews Immunology* 6 (8): 573–83. https://doi.org/10.1038/nri1896.

Oude Munnink, Bas B., Reina S. Sikkema, David F. Nieuwenhuijse, Robert Jan Molenaar, Emmanuelle Munger, Richard Molenkamp, Arco van der Spek, et al. 2021. "Transmission of SARS-CoV-2 on Mink Farms between Humans and Mink and Back to Humans." *Science* 371 (6525): 172–77. https://doi.org/10.1126/science.abe5901.

Peacock, Thomas P., Rebekah Penrice-Randal, Julian A. Hiscox, and Wendy S. Barclay. 2021. "SARS-CoV-2 One Year on: Evidence for Ongoing Viral Adaptation." *Journal of General Virology* 102 (4). https://doi.org/10.1099/jgv.0.001584.

Peck, Kayla M., and Adam S. Lauring. 2018. "Complexities of Viral Mutation Rates." Edited by Christopher S. Sullivan. *Journal of Virology* 92 (14): e01031-17. https://doi.org/10.1128/JVI.01031-17.

Pereira, Filipe. 2020. "Evolutionary Dynamics of the SARS-CoV-2 ORF8 Accessory Gene." *Infection, Genetics and Evolution* 85 (November): 104525. https://doi.org/10.1016/j.meegid.2020.104525.

Petrova, Velislava N., Luke Muir, Paul F. McKay, George S. Vassiliou, Kenneth G. C. Smith, Paul A. Lyons, Colin A. Russell, Carl A. Anderson, Paul Kellam, and Rachael J. M. Bashford-Rogers. 2018. "Combined Influence of B-Cell Receptor Rearrangement and Somatic Hypermutation on B-Cell Class-Switch Fate in Health and in Chronic Lymphocytic Leukemia." *Frontiers in Immunology* 9 (August): 1784. https://doi.org/10.3389/fimmu.2018.01784.

Pettersen, Henrik Sahlin, Anastasia Galashevskaya, Berit Doseth, Mirta M.L. Sousa, Antonio Sarno, Torkild Visnes, Per Arne Aas, et al. 2015. "AID Expression in B-Cell Lymphomas Causes Accumulation of Genomic Uracil and a Distinct AID Mutational Signature." *DNA Repair* 25 (January): 60–71. https://doi.org/10.1016/j.dnarep.2014.11.006.

Plante, Jessica A., Yang Liu, Jianying Liu, Hongjie Xia, Bryan A. Johnson, Kumari G. Lokugamage, Xianwen Zhang, et al. 2021. "Spike Mutation D614G Alters SARS-CoV-2 Fitness." *Nature* 592 (7852): 116–21. https://doi.org/10.1038/s41586-020-2895-3.

Popov, Aleksandr. 2022. "Immunomind/Immunarch: Immunarch 0.7.0." Zenodo.

https://doi.org/10.5281/ZENODO.3367200.

Prado-Vivar, Belén, Mónica Becerra-Wong, Juan José Guadalupe, Sully Márquez, Bernardo Gutierrez, Patricio Rojas-Silva, Michelle Grunauer, Gabriel Trueba, Verónica Barragán, and Paúl Cárdenas. 2021. "A Case of SARS-CoV-2 Reinfection in Ecuador." *The Lancet Infectious Diseases* 21 (6): e142. https://doi.org/10.1016/S1473-3099(20)30910-5.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments." Edited by Art F. Y. Poon. *PLoS ONE* 5 (3): e9490. https://doi.org/10.1371/journal.pone.0009490.

Qian, Guo-Qing, Xue-Qin Chen, Ding-Feng Lv, Ada Hoi Yan Ma, Li-Ping Wang, Nai-Bin Yang, and Xiao-Min Chen. 2020. "Duration of SARS-CoV-2 Viral Shedding during COVID-19 Infection." *Infectious Diseases* 52 (7): 511–12. https://doi.org/10.1080/23744235.2020.1748705.

R Core Team. 2018. "R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,." https://www.R-project.org/.

Ralph, Duncan K., and Frederick A. Matsen. 2020. "Using B Cell Receptor Lineage Structures to Predict Affinity." Edited by Anders Wallqvist. *PLOS Computational Biology* 16 (11): e1008391. https://doi.org/10.1371/journal.pcbi.1008391.

Rambaut, Andrew, Edward C. Holmes, Áine O'Toole, Verity Hill, John T. McCrone, Christopher Ruis, Louis du Plessis, and Oliver G. Pybus. 2020. "A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology." *Nature Microbiology* 5 (11): 1403–7. https://doi.org/10.1038/s41564-020-0770-5.

Reuken, Philipp A., Andreas Stallmach, Mathias W. Pletz, Christian Brandt, Nico Andreas, Sabine Hahnfeld, Bettina Löffler, Sabine Baumgart, Thomas Kamradt, and Michael Bauer. 2021. "Severe Clinical Relapse in an Immunocompromised Host with Persistent SARS-CoV-2 Infection." *Leukemia* 35 (3): 920–23. https://doi.org/10.1038/s41375-021-01175-8.

Reynisson, Birkir, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. 2020. "NetMHCpan-4.1 and NetMHCIIpan-4.0: Improved Predictions of MHC Antigen Presentation by Concurrent Motif Deconvolution and Integration of MS MHC Eluted Ligand Data." *Nucleic Acids Research* 48 (W1): W449–54. https://doi.org/10.1093/nar/gkaa379.

Ria Lassaunière, Jannik Fona, Morten Rasmussen, Anders Frische, Charlotta Polacek Strandh, Thomas Bruun Rasmussen, Anette Bøtner, Anders Fomsgaard ger. 2020. “Working Paper on SARS-CoV-2 Spike Mutations Arising in Danish Mink, Their Spread to Humans and Neutralization Data.” https://files.ssi.dk/Mink-cluster-5-short-report_AFO2.

Robson, Fran, Khadija Shahed Khan, Thi Khanh Le, Clément Paris, Sinem Demirbag, Peter Barfuss, Palma Rocchi, and Wai-Lung Ng. 2020. “Coronavirus RNA Proofreading: Molecular Basis and Therapeutic Targeting.” *Molecular Cell* 79 (5): 710–27. https://doi.org/10.1016/j.molcel.2020.07.027.

Rogozin, Igor B., Artem G. Lada, Alexander Goncearenco, Michael R. Green, Subhajyoti De, German Nudelman, Anna R. Panchenko, Eugene V. Koonin, and Youri I. Pavlov. 2016. “Activation Induced Deaminase Mutational Signature Overlaps with CpG Methylation Sites in Follicular Lymphoma and Other Cancers.” *Scientific Reports* 6 (1): 38133. https://doi.org/10.1038/srep38133.

Rosati, Elisa, C Marie Dowds, Evaggelia Liaskou, Eva Kristine Klemsdal Henriksen, Tom H Karlsen, and Andre Franke. 2017. “Overview of Methodologies for T-Cell Receptor Repertoire Analysis.” *BMC Biotechnology* 17 (1): 61. https://doi.org/10.1186/s12896-017-0379-9.

Roth, David B. 2014. “V(D)J Recombination: Mechanism, Errors, and Fidelity.” Edited by Martin Gellert and Nancy Craig. *Microbiology Spectrum* 2 (6): 2.6.18. https://doi.org/10.1128/microbiolspec.MDNA3-0041-2014.

Sagulenko, Pavel, Vadim Puller, and Richard A Neher. 2018. “TreeTime: Maximum-Likelihood Phylodynamic Analysis.” *Virus Evolution* 4 (1). https://doi.org/10.1093/ve/vex042.

Saini, Sunil Kumar, Ditte Stampe Hersby, Tripti Tamhane, Helle Rus Povlsen, Susana Patricia Amaya Hernandez, Morten Nielsen, Anne Ortved Gang, and Sine Reker Hadrup. 2021. “SARS-CoV-2 Genome-Wide T Cell Epitope Mapping Reveals Immunodominance and Substantial CD8 $^{+}$ T Cell Activation in COVID-19 Patients.” *Science Immunology* 6 (58): eabf7550. https://doi.org/10.1126/sciimmunol.abf7550.

Sarkizova, Siranush, Susan Klaeger, Phuong M. Le, Letitia W. Li, Giacomo Oliveira, Hasmik Keshishian, Christina R. Hartigan, et al. 2020. “A Large Peptidome Dataset Improves HLA Class I Epitope Prediction across Most of the Human Population.” *Nature Biotechnology* 38 (2): 199–209. https://doi.org/10.1038/s41587-019-0322-9.

Schatz, David G., and Yanhong Ji. 2011. “Recombination Centres and the Orchestration of V(D)J Recombination.” *Nature Reviews Immunology* 11 (4): 251–63. https://doi.org/10.1038/nri2941.

Scotto–Lavino, Elizabeth, Guangwei Du, and Michael A Frohman. 2006. “5′ End CDNA Amplification Using Classic RACE.” *Nature Protocols* 1 (6): 2555–62. https://doi.org/10.1038/nprot.2006.480.

Sepulcri, Chiara, Chiara Dentone, Malgorzata Mikulska, Bianca Bruzzone, Alessia Lai, Daniela Fenoglio, Federica Bozzano, et al. 2021. “The Longest Persistence of Viable SARS-CoV-2 with Recurrence of Viremia and Relapsing Symptomatic COVID-19 in an Immunocompromised Patient – a Case Study.” Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2021.01.23.21249554.

Sergei Pond. 2020. “Evolutionary Annotation of Global SARS-CoV-2/COVID-19 Genomes Enabled by Data from GISAID,” December.

Shiehzadegan, Shayan, Nazanin Alaghemand, Michael Fox, and Vishwanath Venketaraman. 2021. “Analysis of the Delta Variant B.1.617.2 COVID-19.” *Clinics and Practice* 11 (4): 778–84. https://doi.org/10.3390/clinpract11040093.

Shu, Yuelong, and John McCauley. 2017. “GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality.” *Eurosurveillance* 22 (13). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

Shugay, Mikhail, Olga V Britanova, Ekaterina M Merzlyak, Maria A Turchaninova, Ilgar Z Mamedov, Timur R Tuganbaev, Dmitriy A Bolotin, et al. 2014. “Towards Error-Free Profiling of Immune Repertoires.” *Nature Methods* 11 (6): 653–55. https://doi.org/10.1038/nmeth.2960.

Simpson, Scott, Fernando U. Kay, Suhny Abbara, Sanjeev Bhalla, Jonathan H. Chung, Michael Chung, Travis S. Henry, et al. 2020. “Radiological Society of North America Expert Consensus Document on Reporting Chest CT Findings Related to COVID-19: Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA.” *Radiology: Cardiothoracic Imaging* 2 (2): e200152. https://doi.org/10.1148/ryct.2020200152.

Sirisinha, Stitaya. 2014. “Evolutionary Insights into the Origin of Innate and Adaptive Immune Systems: Different Shades of Grey.” *Asian Pacific Journal of Allergy and Immunology* 32

(1): 3–15.

Solberg, Owen D., Steven J. Mack, Alex K. Lancaster, Richard M. Single, Yingssu Tsai, Alicia Sanchez-Mazas, and Glenys Thomson. 2008. “Balancing Selection and Heterogeneity across the Classical Human Leukocyte Antigen Loci: A Meta-Analytic Review of 497 Population Studies.” *Human Immunology* 69 (7): 443–64. https://doi.org/10.1016/j.humimm.2008.05.001.

Song, Li, David Cohen, Zhangyi Ouyang, Yang Cao, Xihao Hu, and X. Shirley Liu. 2021. “TRUST4: Immune Repertoire Reconstruction from Bulk and Single-Cell RNA-Seq Data.” *Nature Methods* 18 (6): 627–30. https://doi.org/10.1038/s41592-021-01142-2.

Souza, William M., Mariene R. Amorim, Renata Sesti-Costa, Lais D. Coimbra, Natalia S. Brunetti, Daniel A. Toledo-Teixeira, Gabriela F. de Souza, et al. 2021. “Neutralisation of SARS-CoV-2 Lineage P.1 by Antibodies Elicited through Natural SARS-CoV-2 Infection or Vaccination with an Inactivated SARS-CoV-2 Vaccine: An Immunological Study.” *The Lancet. Microbe* 2 (10): e527–35. https://doi.org/10.1016/S2666-5247(21)00129-4.

Spinelli, Matthew A, Kara L Lynch, Cassandra Yun, David V Glidden, Michael J Peluso, Timothy J Henrich, Monica Gandhi, and Lillian B Brown. 2021. “SARS-CoV-2 Seroprevalence, and IgG Concentration and Pseudovirus Neutralising Antibody Titres after Infection, Compared by HIV Status: A Matched Case-Control Observational Study.” *The Lancet HIV* 8 (6): e334–41. https://doi.org/10.1016/S2352-3018(21)00072-2.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

Starr, Tyler N., Allison J. Greaney, Amin Addetia, William W. Hannon, Manish C. Choudhary, Adam S. Dingens, Jonathan Z. Li, and Jesse D. Bloom. 2021. “Prospective Mapping of Viral Mutations That Escape Antibodies Used to Treat COVID-19.” *Science* 371 (6531): 850–54. https://doi.org/10.1126/science.abf9302.

Stern, Joel N. H., Gur Yaari, Jason A. Vander Heiden, George Church, William F. Donahue, Rogier Q. Hintzen, Anita J. Huttner, et al. 2014. “B Cells Populating the Multiple Sclerosis Brain Mature in the Draining Cervical Lymph Nodes.” *Science Translational Medicine* 6 (248). https://doi.org/10.1126/scitranslmed.3008879.

Tas, Jeroen M. J., Luka Mesin, Giulia Pasqual, Sasha Targ, Johanne T. Jacobsen, Yasuko M.

Mano, Casie S. Chen, et al. 2016. “Visualizing Antibody Affinity Maturation in Germinal Centers.” *Science* 351 (6277): 1048–54. https://doi.org/10.1126/science.aad3439.

Teng, Grace, and F. Nina Papavasiliou. 2007. “Immunoglobulin Somatic Hypermutation.” *Annual Review of Genetics* 41 (1): 107–20. https://doi.org/10.1146/annurev.genet.41.110306.130340.

Tenthorey, Jeannette L., Michael Emerman, and Harmit S. Malik. 2022. “Evolutionary Landscapes of Host-Virus Arms Races.” *Annual Review of Immunology* 40 (1): 271–94. https://doi.org/10.1146/annurev-immunol-072621-084422.

Teraguchi, Shunsuke, Dianita S. Saputri, Mara Anais Llamas-Covarrubias, Ana Davila, Diego Diez, Sedat Aybars Nazlica, John Rozewicki, et al. 2020. “Methods for Sequence and Structural Analysis of B and T Cell Receptor Repertoires.” *Computational and Structural Biotechnology Journal* 18: 2000–2011. https://doi.org/10.1016/j.csbj.2020.07.008.

Tesoriero, James M., Carol-Ann E. Swain, Jennifer L. Pierce, Lucila Zamboni, Meng Wu, David R. Holtgrave, Charles J. Gonzalez, et al. 2021. “COVID-19 Outcomes Among Persons Living With or Without Diagnosed HIV Infection in New York State.” *JAMA Network Open* 4 (2): e2037069. https://doi.org/10.1001/jamanetworkopen.2020.37069.

Tillett, Richard L, Joel R Sevinsky, Paul D Hartley, Heather Kerwin, Natalie Crawford, Andrew Gorzalski, Chris Laverdure, et al. 2021. “Genomic Evidence for Reinfection with SARS-CoV-2: A Case Study.” *The Lancet Infectious Diseases* 21 (1): 52–58. https://doi.org/10.1016/S1473-3099(20)30764-7.

Troyer, Ryan M., John McNevin, Yi Liu, Shao Chong Zhang, Randall W. Krizan, Awet Abraha, Denis M. Tebit, et al. 2009. “Variable Fitness Impact of HIV-1 Escape Mutations to Cytotoxic T Lymphocyte (CTL) Response.” Edited by Christopher M. Walker. *PLoS Pathogens* 5 (4): e1000365. https://doi.org/10.1371/journal.ppat.1000365.

Truong, Thao T., Alex Ryutov, Utsav Pandey, Rebecca Yee, Lior Goldberg, Deepa Bhojwani, Paibel Aguayo-Hiraldo, et al. 2021. “Increased Viral Variants in Children and Young Adults with Impaired Humoral Immunity and Persistent Sars-Cov-2 Infection: A Consecutive Case Series.” *EBioMedicine* 67 (May): 103355. https://doi.org/10.1016/j.ebiom.2021.103355.

Tsueng, Ginger, Julia L. Mullen, Manar Alkuzweny, Marco Cano, Benjamin Rush, Emily Haag, Outbreak Curators, et al. 2022. “Outbreak.Info Research Library: A Standardized, Searchable Platform to Discover and Explore COVID-19 Resources." Preprint. Bioinformatics. https://doi.org/10.1101/2022.01.20.477133.

Turakhia, Yatish, Bryan Thornlow, Angie Hinrichs, Jakob McBroome, Nicolas Ayala, Cheng Ye, Kyle Smith, et al. 2022. "Pandemic-Scale Phylogenomics Reveals The SARS-CoV-2 Recombination Landscape." *Nature*, August. https://doi.org/10.1038/s41586-022-05189-9.

Turchaninova, M A, A Davydov, O V Britanova, M Shugay, V Bikos, E S Egorov, V I Kirgizova, et al. 2016. "High-Quality Full-Length Immunoglobulin Profiling with Unique Molecular Barcoding." *Nature Protocols* 11 (9): 1599–1616. https://doi.org/10.1038/nprot.2016.093.

Tyson, John R, Phillip James, David Stoddart, Natalie Sparks, Arthur Wickenhagen, Grant Hall, Ji Hyun Choi, et al. 2020. "Improvements to the ARTIC Multiplex PCR Method for SARS-CoV-2 Genome Sequencing Using Nanopore." Preprint. Genomics. https://doi.org/10.1101/2020.09.04.283077.

Valpione, Sara, Piyushkumar A. Mundra, Elena Galvani, Luca G. Campana, Paul Lorigan, Francesco De Rosa, Avinash Gupta, et al. 2021. "The T Cell Receptor Repertoire of Tumor Infiltrating T Cells Is Predictive and Prognostic for Cancer Survival." *Nature Communications* 12 (1): 4098. https://doi.org/10.1038/s41467-021-24343-x.

Vander Heiden, Jason A., Gur Yaari, Mohamed Uduman, Joel N.H. Stern, Kevin C. O'Connor, David A. Hafler, Francois Vigneault, and Steven H. Kleinstein. 2014. "PRESTO: A Toolkit for Processing High-Throughput Sequencing Raw Reads of Lymphocyte Receptor Repertoires." *Bioinformatics* 30 (13): 1930–32. https://doi.org/10.1093/bioinformatics/btu138.

Vita, Randi, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. 2019. "The Immune Epitope Database (IEDB): 2018 Update." *Nucleic Acids Research* 47 (D1): D339–43. https://doi.org/10.1093/nar/gky1006.

V'kovski, Philip, Annika Kratzel, Silvio Steiner, Hanspeter Stalder, and Volker Thiel. 2021. "Coronavirus Biology and Replication: Implications for SARS-CoV-2." *Nature Reviews. Microbiology* 19 (3): 155–70. https://doi.org/10.1038/s41579-020-00468-6.

Voloch, Carolina M, Ronaldo da Silva Francisco Jr, Luiz G P de Almeida, Otavio J Brustolini, Cynthia C Cardoso, Alexandra L Gerber, Ana Paula de C Guimarães, et al. 2021. “Intra-Host Evolution during SARS-CoV-2 Prolonged Infection.” *Virus Evolution* 7 (2): veab078. https://doi.org/10.1093/ve/veab078.

Walls, Alexandra C., Young-Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, and David Veesler. 2020. “Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.” *Cell* 181 (2): 281-292.e6. https://doi.org/10.1016/j.cell.2020.02.058.

Wang. 2021. “Safety and Efficacy of the BNT162b2 MRNA Covid-19 Vaccine.” *New England Journal of Medicine* 384 (16): 1576–78. https://doi.org/10.1056/NEJMc2036242.

Wang, Rui, Jiahui Chen, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. 2020. “Decoding Asymptomatic COVID-19 Infection and Transmission.” *The Journal of Physical Chemistry Letters* 11 (23): 10007–15. https://doi.org/10.1021/acs.jpclett.0c02765.

Watson, Lisa C., Chantelle S. Moffatt-Blue, R. Zachary McDonald, Elizabeth Kompfner, Djemel Ait-Azzouzene, David Nemazee, Argyrios N. Theofilopoulos, Dwight H. Kono, and Ann J. Feeney. 2006. “Paucity of V-D-D-J Rearrangements and V $_{H}$ Replacement Events in Lupus Prone and Nonautoimmune TdT $^{-/-}$ and TdT $^{+/+}$ Mice.” *The Journal of Immunology* 177 (2): 1120–28. https://doi.org/10.4049/jimmunol.177.2.1120.

Wei, Lai, Bin Liu, Yuanyuan Zhao, and Zhishui Chen. 2020. “Prolonged Shedding of SARS-CoV-2 in an Elderly Liver Transplant Patient Infected by COVID-19: A Case Report.” *Annals of Palliative Medicine* 9 (5): 8–8. https://doi.org/10.21037/apm-20-996.

Wei, Xiping, Julie M. Decker, Shuyi Wang, Huxiong Hui, John C. Kappes, Xiaoyun Wu, Jesus F. Salazar-Gonzalez, et al. 2003. “Antibody Neutralization and Escape by HIV-1.” *Nature* 422 (6929): 307–12. https://doi.org/10.1038/nature01470.

Weigang, Sebastian, Jonas Fuchs, Gert Zimmer, Daniel Schnepf, Lisa Kern, Julius Beer, Hendrik Luxenburger, et al. 2021. “Within-Host Evolution of SARS-CoV-2 in an Immunosuppressed COVID-19 Patient as a Source of Immune Escape Variants.” *Nature Communications* 12 (1): 6405. https://doi.org/10.1038/s41467-021-26602-3.

Weisblum, Yiska, Fabian Schmidt, Fengwen Zhang, Justin DaSilva, Daniel Poston, Julio CC Lorenzi, Frauke Muecksch, et al. 2020. “Escape from Neutralizing Antibodies by SARS-CoV-2 Spike Protein Variants.” *ELife* 9 (October): e61312. https://doi.org/10.7554/eLife.61312.

Weitz, Joshua S, Guanlin Li, Hayriye Gulbudak, Michael H Cortez, and Rachel J Whitaker. 2019. “Viral Invasion Fitness across a Continuum from Lysis to Latency†.” *Virus Evolution* 5 (1): vez006. https://doi.org/10.1093/ve/vez006.

Wertheim, Joel O., Jade C. Wang, Mindy Leelawong, Darren P. Martin, Jennifer L. Havens, Moinuddin A. Chowdhury, Jonathan E. Pekar, et al. 2022. “Detection of SARS-CoV-2 Intra-Host Recombination during Superinfection with Alpha and Epsilon Variants in New York City.” *Nature Communications* 13 (1): 3645. https://doi.org/10.1038/s41467-022-31247-x.

WHO. 2022. “Tracking SARS-CoV-2 Variants. [Cited 25 Aug 2022]. Available: Https://Www.Who.Int/Activities/Tracking-SARS-CoV-2-Variants.” 2022.

Williamson, Maia Kavanagh, Fergus Hamilton, Stephanie Hutchings, Hannah M. Pymont, Mark Hackett, David Arnold, Nick A Maskell, et al. 2021. “Chronic SARS-CoV-2 Infection and Viral Evolution in a Hypogammaglobulinaemic Individual.” Preprint. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2021.05.31.21257591.

Wilm, Andreas, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. 2012. “LoFreq: A Sequence-Quality Aware, Ultra-Sensitive Variant Caller for Uncovering Cell-Population Heterogeneity from High-Throughput Sequencing Datasets.” *Nucleic Acids Research* 40 (22): 11189–201. https://doi.org/10.1093/nar/gks918.

Wölfel, Roman, Victor M. Corman, Wolfgang Guggemos, Michael Seilmaier, Sabine Zange, Marcel A. Müller, Daniela Niemeyer, et al. 2020. “Virological Assessment of Hospitalized Patients with COVID-2019.” *Nature* 581 (7809): 465–69. https://doi.org/10.1038/s41586-020-2196-x.

Wong, Phillip, and Eric G. Pamer. 2003. “CD8 T Cell Responses to Infectious Pathogens.” *Annual Review of Immunology* 21 (1): 29–70. https://doi.org/10.1146/annurev.immunol.21.120601.141114.

Wu, Fan, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, et al. 2020. “A New Coronavirus Associated with Human Respiratory Disease in China.” *Nature* 579 (7798): 265–69. https://doi.org/10.1038/s41586-020-2008-3.

Xu, Yuexin, Alicia J. Morales, Andrea M. H. Towlerton, Shreeram Akilesh, Chris P. Miller, Scott

S. Tykodi, and Edus H. Warren. 2022. “Integrated TCR Repertoire Analysis and Single-Cell Transcriptomic Profiling of Tumor-Infiltrating T Cells in Renal Cell Carcinoma Identifies Shared and Tumor-Restricted Expanded Clones with Unique Phenotypes.” *Frontiers in Oncology* 12 (September): 952252. https://doi.org/10.3389/fonc.2022.952252.

Yaari, Gur, Jason A. Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Namita Gupta, Joel N. H. Stern, Kevin C. O’Connor, et al. 2013. “Models of Somatic Hypermutation Targeting and Substitution Based on Synonymous Mutations from High-Throughput Immunoglobulin Sequencing Data.” *Frontiers in Immunology* 4. https://doi.org/10.3389/fimmu.2013.00358.

Yang, Ziheng. 2006. *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford: Oxford University Press.

Yermanos, Alexander, Victor Greiff, Tanja Stadler, Annette Oxenius, and Sai T. Reddy. 2020. “The Influence of the Phylogenetic Inference Pipeline on Murine Antibody Repertoire Sequencing Data Following Viral Infection.” Preprint. Bioinformatics. https://doi.org/10.1101/2020.03.20.000521.

Yi, Kijong, Su Yeon Kim, Thomas Bleazard, Taewoo Kim, Jeonghwan Youk, and Young Seok Ju. 2021. “Mutational Spectrum of SARS-CoV-2 during the Global Pandemic.” *Experimental & Molecular Medicine* 53 (8): 1229–37. https://doi.org/10.1038/s12276-021-00658-z.

Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan‐Yuk Lam. 2017. “Ggtree: Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data.” Edited by Greg McInerny. *Methods in Ecology and Evolution* 8 (1): 28–36. https://doi.org/10.1111/2041-210X.12628.

Zabalza, Ana, Simón Cárdenas‐Robledo, Paula Tagliani, Georgina Arrambide, Susana Otero‐Romero, Pere Carbonell‐Mirabent, Marta Rodriguez‐Barranco, et al. 2021. “COVID‐19 in Multiple Sclerosis Patients: Susceptibility, Severity Risk Factors and Serological Response.” *European Journal of Neurology* 28 (10): 3384–95. https://doi.org/10.1111/ene.14690.

Zhang, Qing, Christian M. Zmasek, and Adam Godzik. 2010. “Domain Architecture Evolution of Pattern-Recognition Receptors.” *Immunogenetics* 62 (5): 263–72.

https://doi.org/10.1007/s00251-010-0428-1.

Zhang, Yiwen, Yingshi Chen, Yuzhuang Li, Feng Huang, Baohong Luo, Yaochang Yuan, Baijin Xia, et al. 2021. “The ORF8 Protein of SARS-CoV-2 Mediates Immune Evasion through down-Regulating MHC-I.” *Proceedings of the National Academy of Sciences* 118 (23): e2024202118. https://doi.org/10.1073/pnas.2024202118.

Zhou, Jeffrey O., Hussain A. Zaidi, Therese Ton, and Daniela Fera. 2020. “The Effects of Framework Mutations at the Variable Domain Interface on Antibody Affinity Maturation in an HIV-1 Broadly Neutralizing Antibody Lineage.” *Frontiers in Immunology* 11 (July): 1529. https://doi.org/10.3389/fimmu.2020.01529.

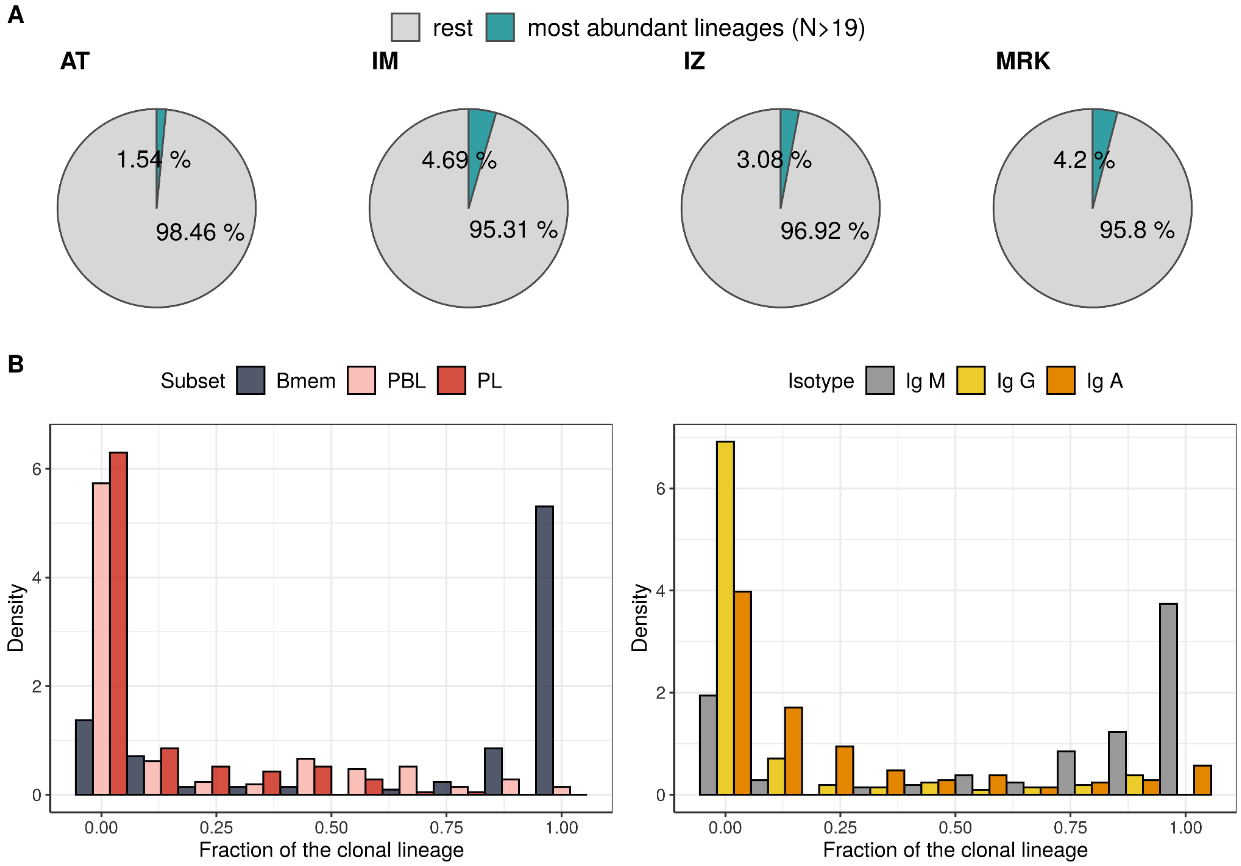
Zhu, Jinfang, Hidehiro Yamane, and William E. Paul. 2010. “Differentiation of Effector CD4 T Cell Populations.” *Annual Review of Immunology* 28 (1): 445–89. https://doi.org/10.1146/annurev-immunol-030409-101212.

Zinzula, Luca. 2021. “Lost in Deletion: The Enigmatic ORF8 Protein of SARS-CoV-2.” *Biochemical and Biophysical Research Communications* 538 (January): 116–24. https://doi.org/10.1016/j.bbrc.2020.10.045.

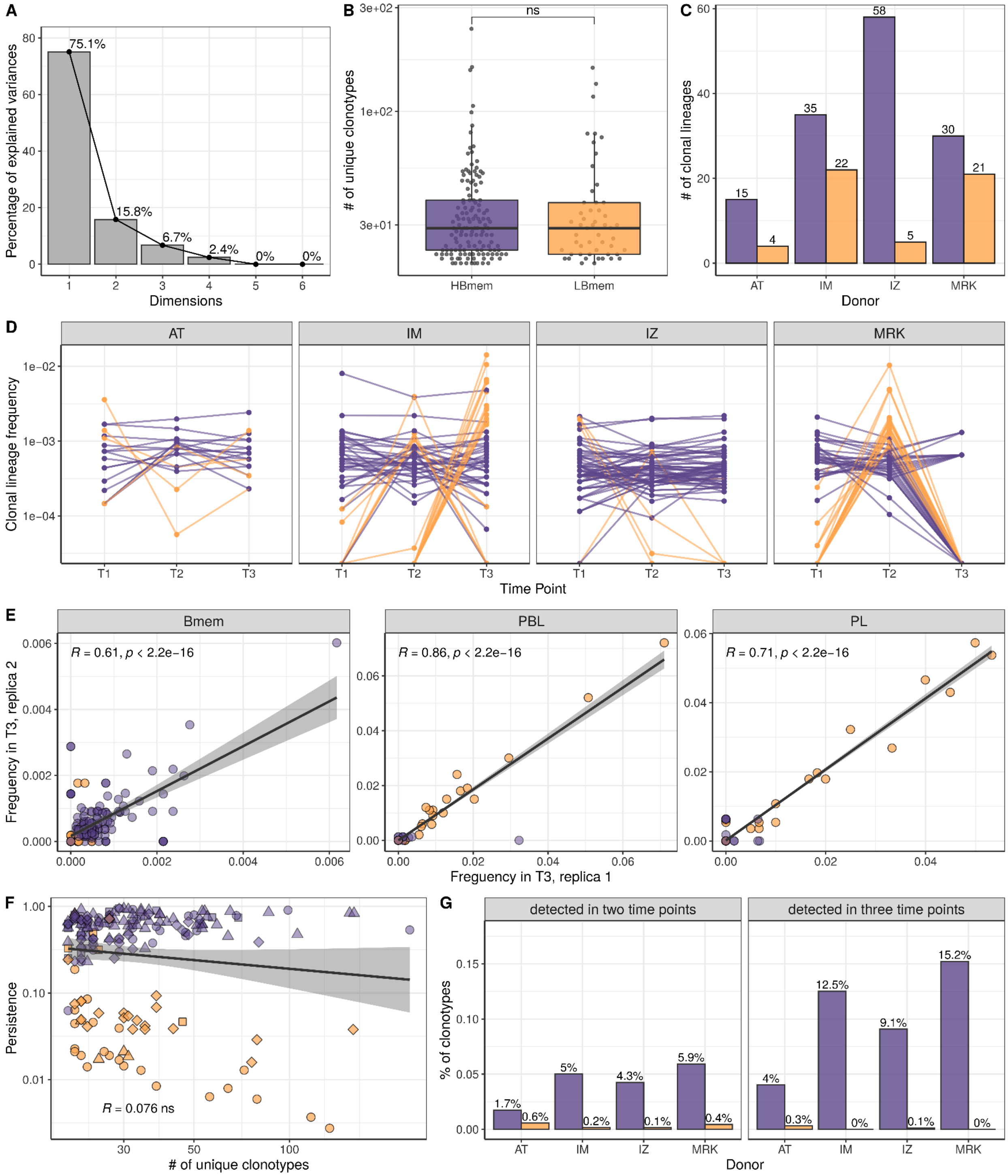
Zumla, Alimuddin, David S Hui, and Stanley Perlman. 2015. “Middle East Respiratory Syndrome.” *The Lancet* 386 (9997): 995–1007. https://doi.org/10.1016/S0140-6736(15)60454-8.

# APPENDIX A

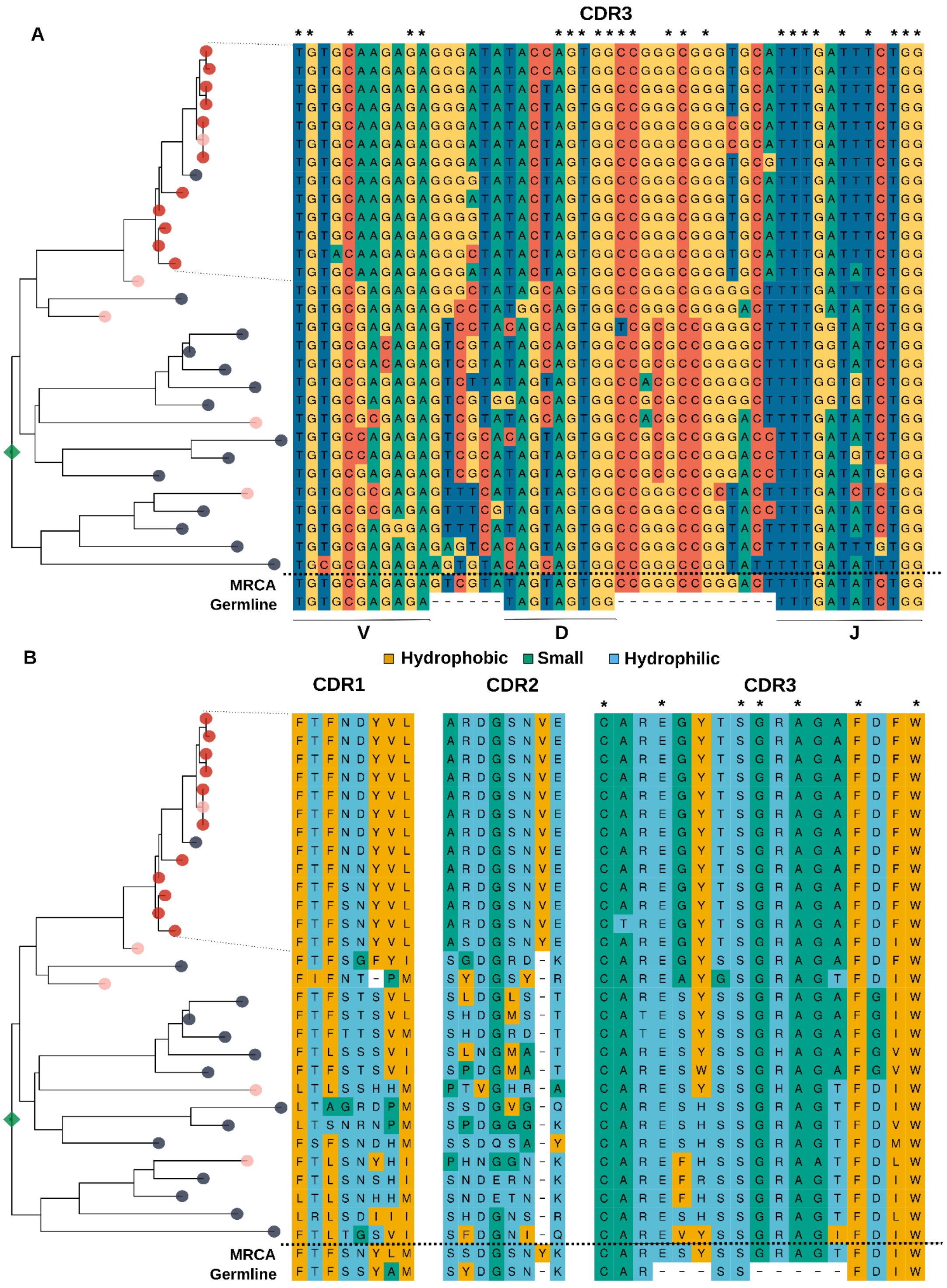
## Supplementary Figures A



**Supplementary Figure A-1. A**: Proportion of IGH clonotype diversity occupied by the most abundant clonal lineages (⩾ 20 unique clonotypes). **B**: Distributions of fractions of cellular subtypes and isotypes in most abundant clonal lineages.



**Supplementary Figure A-2. A**: Scree plot for the principal component analysis (PCA) from **Figure 3.2A** of the composition of clonal lineages, where fractions of Bmem, PBL, PL, and fractions of IgM, IgG and IgA were used as variables; **B**: Distribution of the number of unique clonotypes in a lineage for HBmem and LBmem; **C**: The number of clonal lineages belonging to HBmem or LBmem clusters in each donor; **D**: Dynamics of clonal lineage frequency from **Figure 3.2C**, presented for individual donors. Lineage frequency is defined as the number of clonotypes in a lineage divided by the total number of clonotypes detected at a given time-point. Each line connects points representing a unique clonal lineage; **E**: Spearman's correlation between frequencies of clonal lineages in two replicates of time-point 3 (T3) samples. Only clonal lineages sampled with at least one replica at this time-point were included in the analysis; **F**: Spearman's correlation between the size of a clonal lineage and its persistence. **G**: Fraction of clonotypes in HBmem or LBmem clonal lineages detected at two or three time-points.



**Supplementary Figure A-3.** Phylogenetic tree and nucleotides (**A**) and amino-acid (**B**) alignments of CDR regions of clonal lineage with example of HBmem-LBmem transition from **Figure 3.3F**. The order of rows in the alignment corresponds to the order of clonotypes on the phylogenetic tree. Rows of the alignment, corresponding to the LBmem - like clade indicated by dotted lines. Colors of cells on figure B represent physicochemical properties of corresponding amino acids. Asterisks indicate positions, conservative among all clonotypes of the lineage.

## Supplementary Tables A

**Supplementary Table A-1. Donor demographics and cell sample sizes.** Multiple values in a cell separated by a semicolon represent replicates collected for the corresponding donor, time point, or cellular subset. AR - allergic rhinitis; FA - food allergy; HD - healthy donor. The table is adapted from (Mikelov et al., 2022).

| | | | | **Number of cells per sample** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Time point | | | T1 | | | T2 | | | T3 | |
| Donor ID | Age | Sex | Status | Bmem | PBL | PL | Bmem | PBL | PL | Bmem | PBL | PL |
| D01 | 27 | F | AR | n/a | n/a | n/a | 50,300;<br>55,400 | 2,100;<br>2,100 | 1,020;<br>1,010 | 50,000;<br>50,000 | 1,000;<br>1,000 | 500;<br>500 |
| IM | 39 | M | AR,FA | 186,572 | 2,200 | 129 | 69,900;<br>68,400 | 2,000;<br>2,486 | 920 | 50,000;<br>50,000 | 2,000;<br>2,000 | 1,000;<br>1,000 |
| MRK | 27 | M | AR | 143,162 | 5,336 | 251 | 51,700;<br>50,600 | 2,130;<br>2,020 | 1,000;<br>1,035 | 50,000;<br>50,000 | 1,000;<br>1,000 | 400;<br>200 |
| AT | 23 | M | AR,FA | 101,400 | 7,200 | 1,800 | 50,600;<br>57,400 | 2,520 | 800 | 50000;<br>40800 | 1,000;<br>1,000 | 400;<br>200 |
| IZ | 33 | M | HD | 101,800 | 3,900 | 850 | 50,500;<br>56,300 | 1,140;<br>1,840 | 1,050;<br>625 | 50,000;<br>50,000 | 2,000;<br>2,000 | 200;<br>200 |
| MT | 33 | F | HD | n/a | n/a | n/a | n/a | n/a | n/a | 50,000;<br>50,000 | 1,000;<br>1,000 | 400 |

**Supplementary Table A-2.** Examples of divergent (*D*) and polymorphic (*P*) sites as calculated for the McDonald-Kreitman (MK) test. Synonymous and nonsynonymous substitutions from germ-line sequence are underlined and boldfaced, correspondingly.

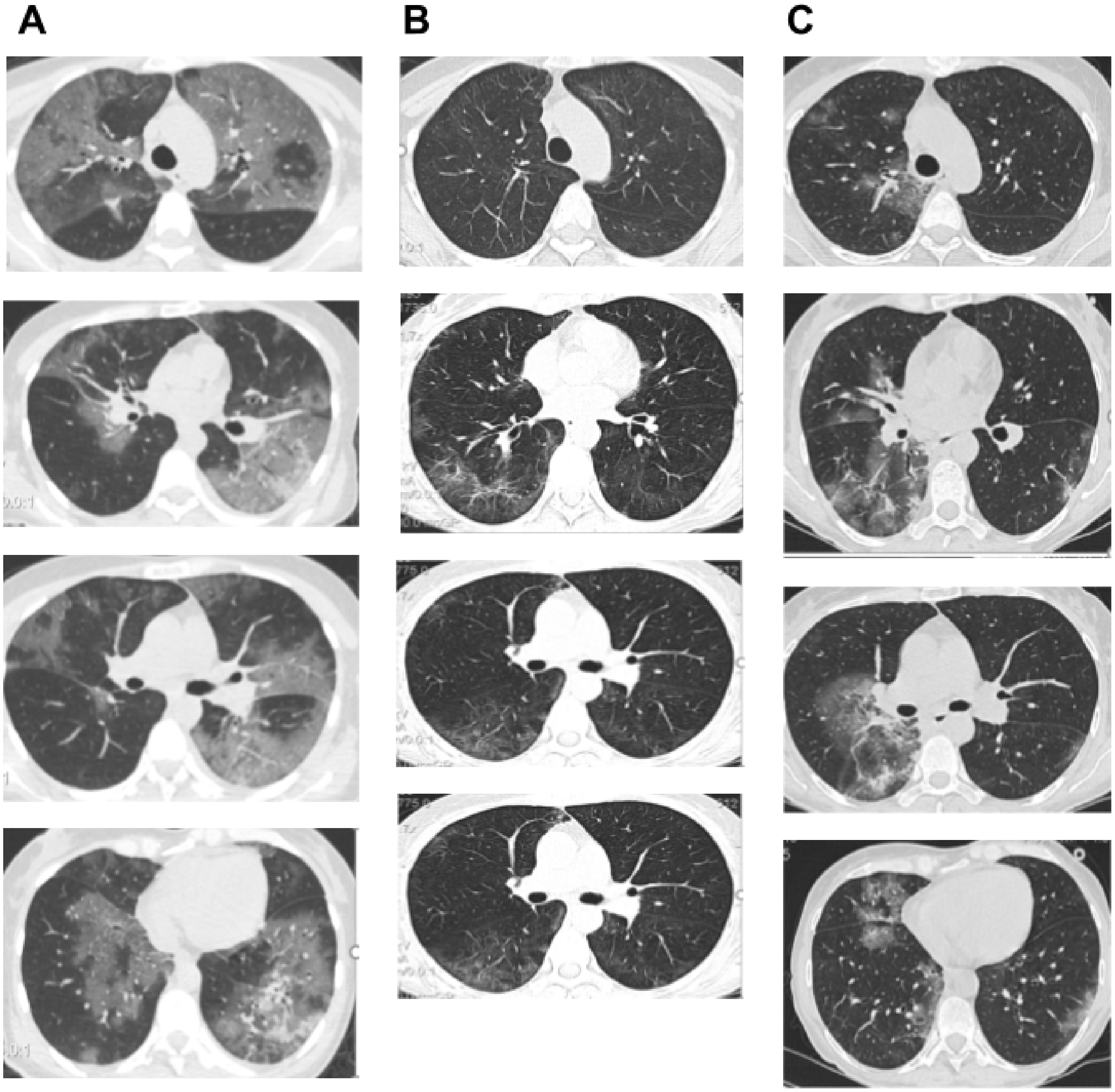
| **Codon #** | **i** | | | **j** | | | **q** | | | **r** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Germline** | c | g | c | g | - | - | c | t | a | a | a | t |
| **MRCA** | c | g | t | g | t | a | c | t | c | a | **g** | t |
| **Clonotypes in the clonal lineage** | c<br>c<br>c<br>c | g<br>**t**<br>g<br>**a** | t<br>t<br>t<br>t | g<br>g<br>g<br>g | t<br>t<br>t<br>t | a<br>a<br>a<br>a | c<br>c<br>c<br>c | t<br>t<br>t<br>t | c<br>a<br>c<br>c | a<br>a<br>a<br>a | **g**<br>**g**<br>**c**<br>**c** | t<br>t<br>t<br>t |
| ***Nonsynonymous divergence (Dn)*** | 0 | 0 | 0 | - | - | - | 0 | 0 | 0 | 0 | **1** | 0 |
| ***Synonymous divergence (Ds)*** | 0 | 0 | 1 | - | - | - | 0 | 0 | 0 | 0 | 0 | 0 |
| ***Nonsynonymous polymorphism (Pn)*** | 0 | **2** | 0 | - | - | - | 0 | 0 | 0 | 0 | **1** | 0 |
| ***Synonymous polymorphism (Ps)*** | 0 | 0 | 0 | - | - | - | 0 | 0 | 1 | 0 | 0 | 0 |
| **Comment** | example of codon represented by multiple variants in the clonal lineage (*i.e.* with the multiallelic site) | | | example of codon excluded from analysis because of unknown germline sequence for the site | | | example of codon where divergence is not counted because of presence of the germline variant among sequence variants in the lineage | | | example of codon where divergence is counted because there are no clonotypes identical to the germline sequence | | |

**Supplementary Table A-3.** MK test results under different inclusion criterion for clonal lineages from HBmem and LBmem clusters, which makes it possible to deal with zero values in G-MRCA nonsynonymous or synonymous divergence. The LBmem cluster demonstrated consistent results of the MK test under all inclusion criteria, and the α of joined inside cluster divergence (combined SHM for all lineages belonging to the cluster) corresponds well to the median α among clonal lineages. The HBmem cluster is better suited for this type of filter, since it generally has much lower G-MRCA distance, and some clonal lineages have no divergence in MRCA from reconstructed portions of the germline sequence. Estimated α on joined cluster divergence in the HBmem cluster varies depending on the type of the filter employed, but is always lower than the α of LBmem. Additionally, consideration of all clonal lineages with addition of pseudocounts to *Dn* and *Ds* produces a negative median α, because the α of a clonal lineage with zero G-MRCA distance will always produce a negative α.

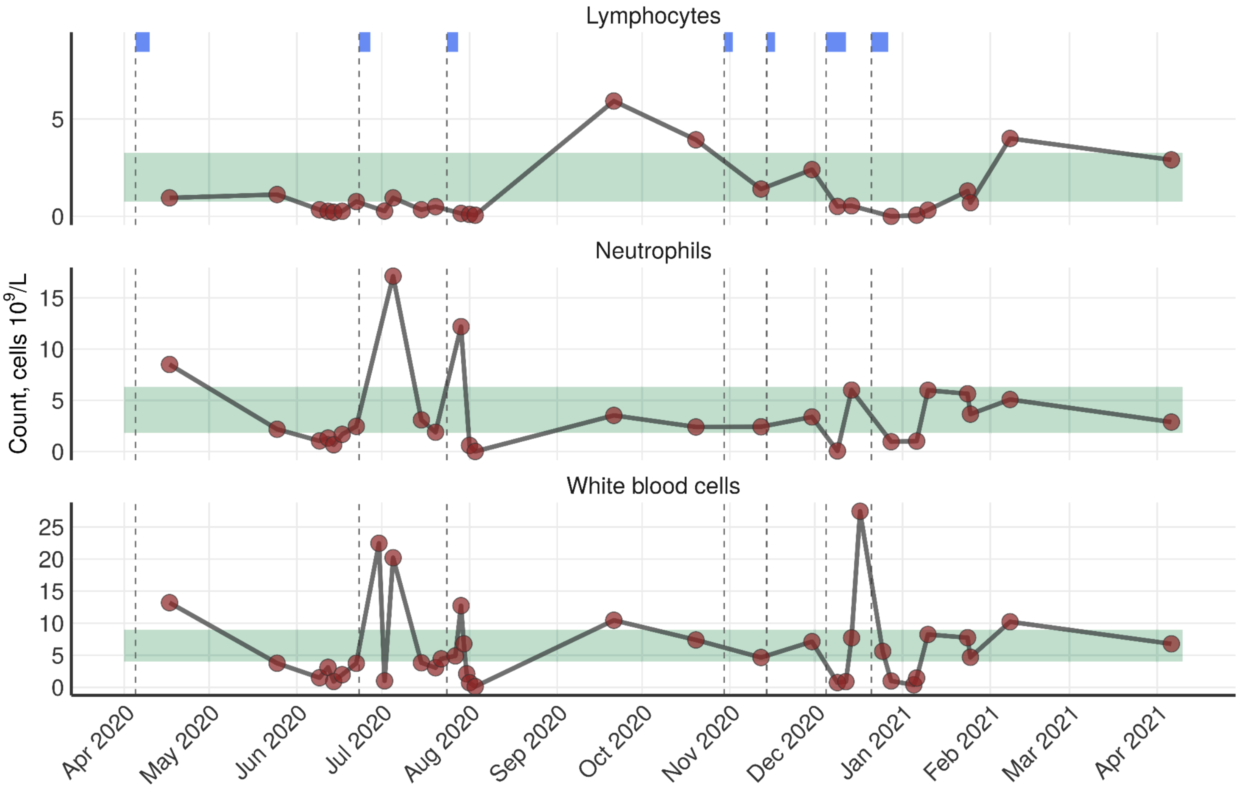
| **Inclusion criteria** | All clonal lineages. Pseudocounts are added to *Dn* and *Ds* to deal with zero values in the MK test of distinct clonal lineages. | | Clonal lineages with nonzero G-MRCA distance (at least one nonsynonymous or synonymous substitution). Pseudocounts are added to *Dn* and *Ds* to deal with zero values in the MK test of distinct clonal lineages. | | Clonal lineages with at least one nonsynonymous and synonymous substitution. No pseudocounts in *Dn* and *Ds* are required. | |
|---|---|---|---|---|---|---|
| **Cluster** | HBmem | LBmem | HBmem | LBmem | HBmem | LBmem |
| **# of filtered clonal lineages** | 138 | 52 | 68 | 49 | 18 | 29 |
| **Median α** | -0.46 | 0.55 | 0.18 | 0.57 | - 0.07 | 0.54 |
| **Mann-Whitney test** | $p = 2.9\cdot10^{-11}$ | | $p = 4.8\cdot10^{-6}$ | | $p = 0.0028$ | |
| **MK test on joined diversity of the cluster** | α = 0.58<br>$p = 4.97\cdot10^{-7}$ | α = 0.65<br>$p < 2.2\cdot10^{-16}$ | α = 0.61<br>$p = 6.05\cdot10^{-8}$ | α = 0.66<br>$p < 2.2\cdot10^{-16}$ | α = 0.26<br>$p = 0.1004$ | α = 0.56<br>$p = 2.05\cdot10^{-10}$ |

# APPENDIX B

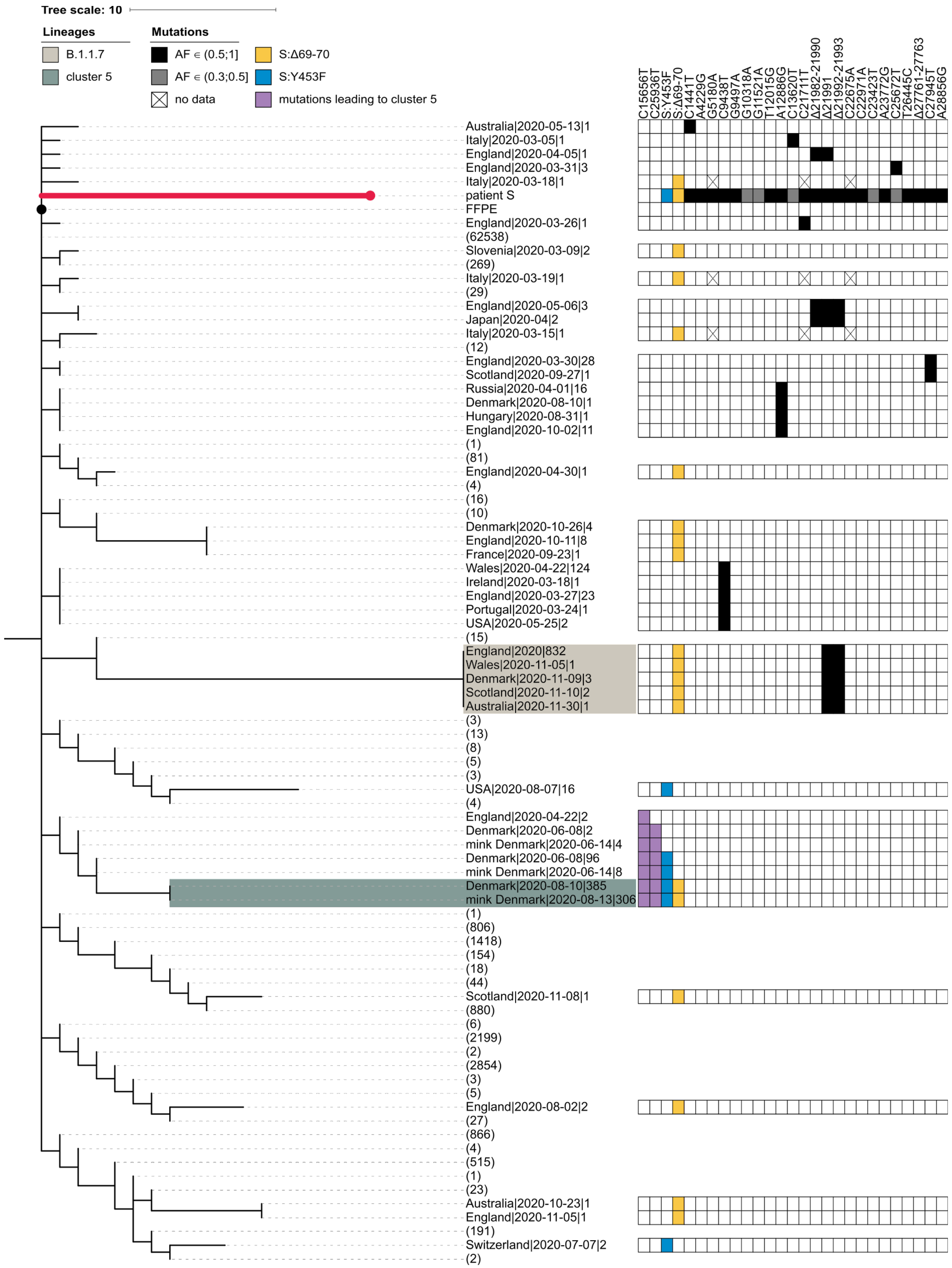
## Supplementary Figures B



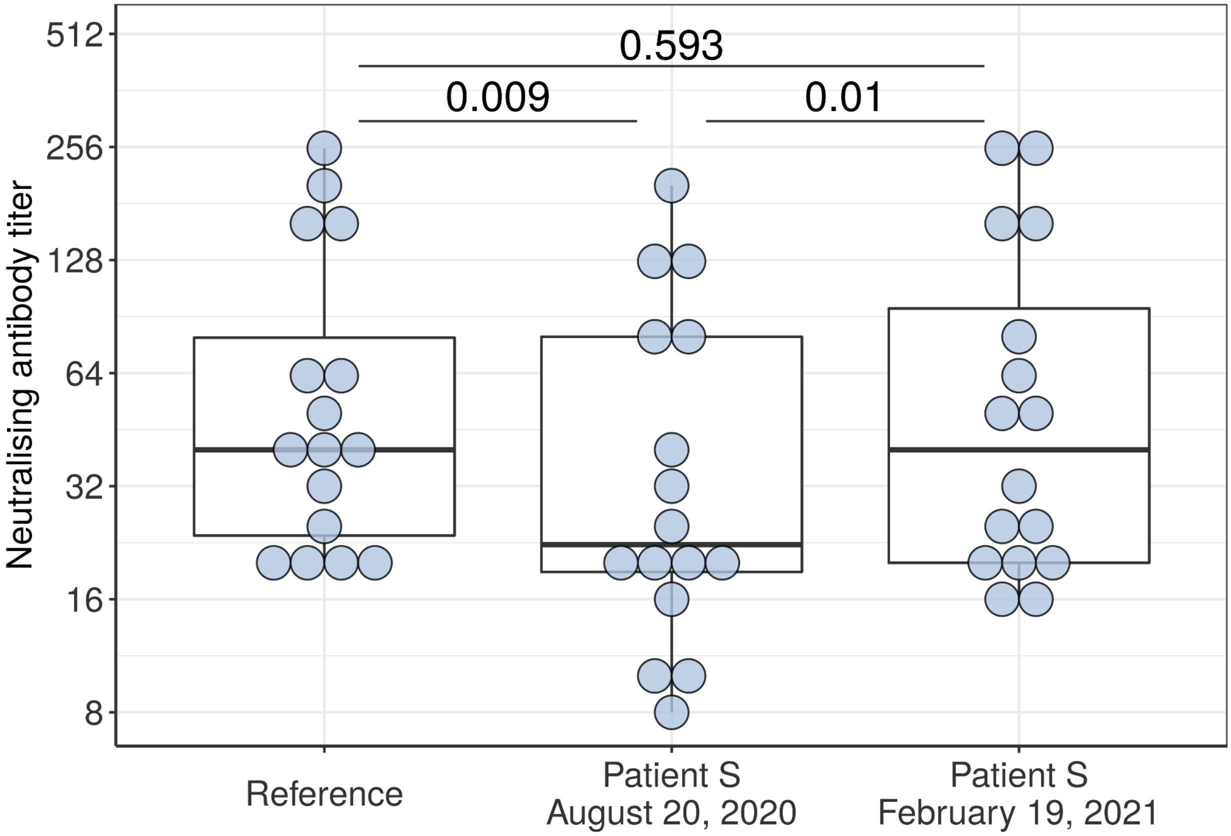
**Supplementary Figure B-1.** CT scans of patient S lungs obtained by Optima 660 (GE) CT-scanner with standard in-house protocols without contrast enhancement in June 2020 (**A**), August 2020 (**B**) and January 2021 (**C**). In both lungs in **(A), (B), (C)** there were bilateral, multifocal and diffuse ground-glass opacifications with small regions of subpleural consolidations, but without predominant distribution in **(A)**, mild reticulation and regions of architectural distortion with the formation of subpleural bands in **(B)**, small regions of mild reticulation, vascular dilatation, and regions of  linear consolidation with the formation of bands in **(C)**. CT-patterns can be determined as typical for COVID-19 pulmonary disease according to the Radiological Society of North America expert consensus (Simpson et al. 2020).  The figure is adapted from (Stanevich et al., 2023).



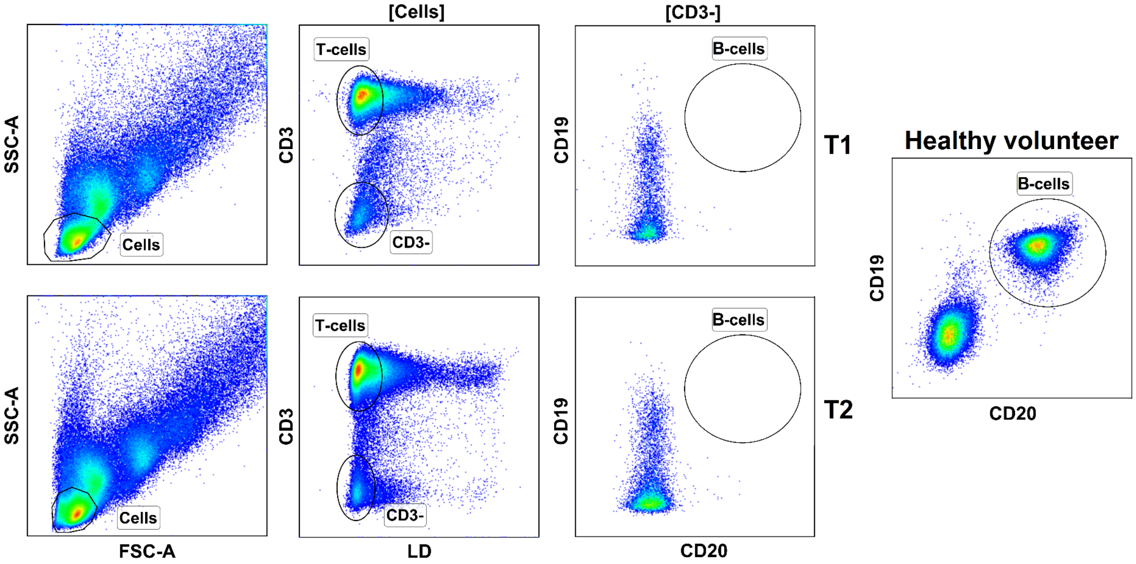
**Supplementary Figure B-2**. Cell counts of lymphocytes, neutrophils and white blood cells in $10^9$/L observed in patient S between April 17, 2020 and June 4, 2021. Green shading indicates the normal range for females of the corresponding age in each category respectively. Blue bars and dashed lines indicate courses of chemotherapy; chemotherapy regimens are described in Supplementary Table B-1 and Figure 4.1A.


Tree scale: 10
Lineages
B.1.1.7
cluster 5
Mutations
AF ∈ (0.5;1]
AF ∈ (0.3;0.5]
no data
S:Δ69-70
S:Y453F
mutations leading to cluster 5
C15656T
C25936T
S:Y453F
S:Δ69-70
C1441T
A4229G
G5180A
C9438T
G9497A
G10318A
G11521A
T12015G
A12886G
C13620T
C21711T
Δ21982-21990
Δ21991
Δ21992-21993
C22675A
C22971A
C23423T
A23772G
C25672T
T26445C
Δ27761-27763
C27945T
A28856G
Australia|2020-05-13|1
Italy|2020-03-05|1
England|2020-04-05|1
England|2020-03-31|3
Italy|2020-03-18|1
patient S
FFPE
England|2020-03-26|1
(62538)
Slovenia|2020-03-09|2
(269)
Italy|2020-03-19|1
(29)
England|2020-05-06|3
Japan|2020-04|2
Italy|2020-03-15|1
(12)
England|2020-03-30|28
Scotland|2020-09-27|1
Russia|2020-04-01|16
Denmark|2020-08-10|1
Hungary|2020-08-31|1
England|2020-10-02|11
(1)
(81)
England|2020-04-30|1
(4)
(16)
(10)
Denmark|2020-10-26|4
England|2020-10-11|8
France|2020-09-23|1
Wales|2020-04-22|124
Ireland|2020-03-18|1
England|2020-03-27|23
Portugal|2020-03-24|1
USA|2020-05-25|2
(15)
England|2020|832
Wales|2020-11-05|1
Denmark|2020-11-09|3
Scotland|2020-11-10|2
Australia|2020-11-30|1
(3)
(13)
(8)
(5)
(3)
USA|2020-08-07|16
(4)
England|2020-04-22|2
Denmark|2020-06-08|2
mink Denmark|2020-06-14|4
Denmark|2020-06-08|96
mink Denmark|2020-06-14|8
Denmark|2020-08-10|385
mink Denmark|2020-08-13|306
(1)
(806)
(1418)
(154)
(18)
(44)
Scotland|2020-11-08|1
(880)
(6)
(2199)
(2)
(2854)
(3)
(5)
England|2020-08-02|2
(27)
(866)
(4)
(515)
(1)
(23)
Australia|2020-10-23|1
England|2020-11-05|1
(191)
Switzerland|2020-07-07|2
(2)

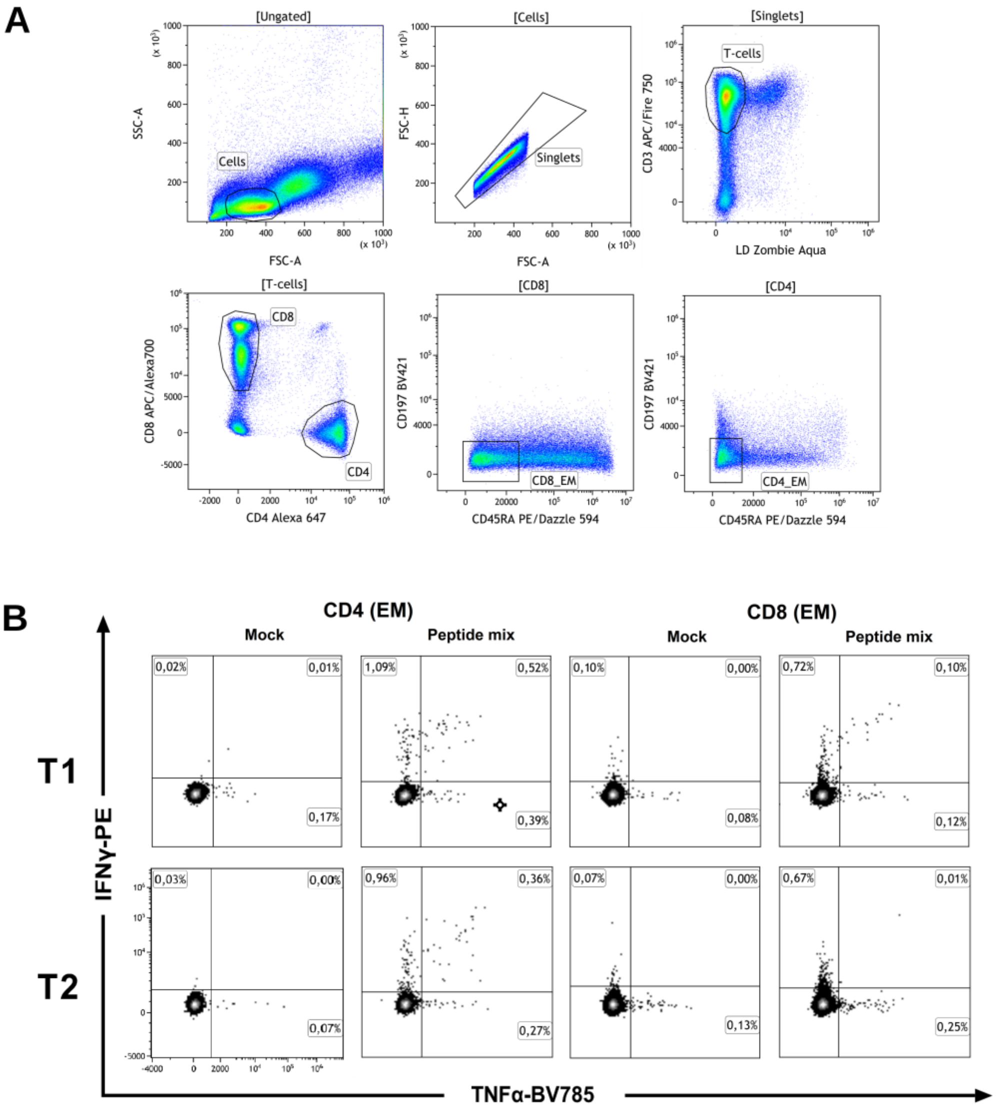
**Supplementary Figure B-3. Patient S is robustly placed outside the cluster 5 clade.** The abridged phylogeny of the B.1.1 lineage phylogeny is shown. Only those samples are shown which met either of the following conditions: (i) carried any of the differences found between the B.1.1 root and the patient S sample (black cells), and these mutations had occurred in the branch immediately descendant from the B.1.1 root; or (ii) carried either the S:Δ69-70HV (blue cells) or the S:Y435F (yellow cells) mutation, independent of the timing of their origin. Additionally, we retained the samples from the branches that separate the cluster 5 clade from the rest of the phylogeny (two additional mutations, purple cells). Samples that didn't meet these criteria were collapsed, with the number of such samples shown in parentheses. The retained samples were then grouped by country, with names formatted as 'country|date of the earliest sample|number of samples'. B.1.1.7 and cluster 5 samples are shaded as in Figure 4.1B. The presence of the above-mentioned mutations is indicated by the matrix at the right. Two mutations distinguishing cluster 5 from the B.1.1 root (purple) reject uniting patient S and cluster 5 in the same clade. For patient S, mutations with allele frequency below 50% in all three samples are shown in grey. Missing data ('N's in sequences) are shown as crosses. FFPE (black dot), patient A sample (the presumed source of infection for patient S). The figure is adapted from (Stanevich et al., 2023).

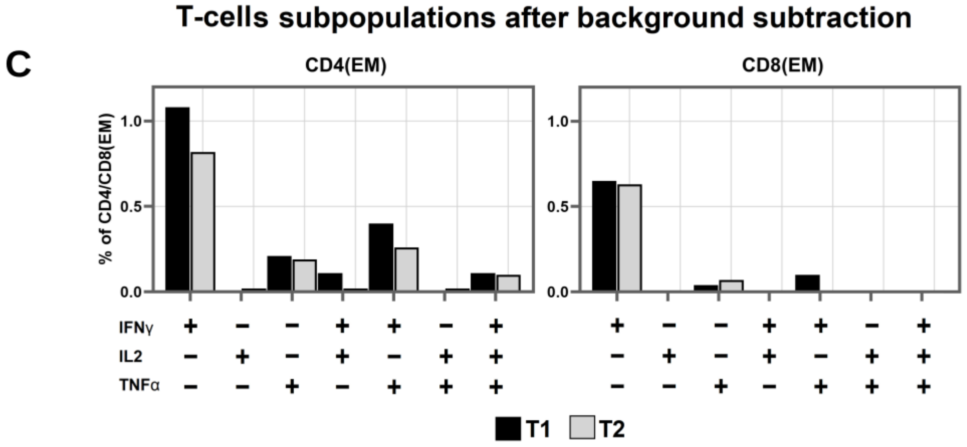


**Supplementary Figure B-4. Effect of viral evolution in patient S on neutralization by antibodies.** Neutralizing activity of serum obtained from 16 convalescent donors against patient S virus samples obtained on August 20, 2020 and February 19, 2021, as well as a reference viral strain of the B.1 lineage isolated from a swab sample in the beginning of the pandemic in Russia in March 2020. The August 20, 2020 isolate demonstrated reduced sensitivity to neutralizing antibodies compared to the reference strain, with the geometric mean fold decrease of 1.6 (CI 1.2-2.0, range 0.8-4.0). The February 19, 2021 isolate carried no signature of reduced sensitivity, and was indistinguishable in its sensitivity to neutralizing antibodies from the reference strain. Each sample was tested in triplicate and GMTs are plotted. Mann-Whitney-Wilcoxon test with Holm adjustments was used for pairwise comparisons. The figure is adapted from (Stanevich et al., 2023).



**Supplementary Figure B-5.** Flow cytometry plots showing the absence of B cells. The figure is adapted from (Stanevich et al., 2023).


A
[Ungated]
SSC-A
FSC-A
Cells
[Cells]
FSC-H
Singlets
[Singlets]
CD3 APC/Fire 750
LD Zombie Aqua
T-cells
[T-cells]
CD8 APC/Alexa700
CD4 Alexa 647
CD8
CD4
[CD8]
CD197 BV421
CD45RA PE/Dazzle 594
CD8_EM
[CD4]
CD197 BV421
CD45RA PE/Dazzle 594
CD4_EM
B
CD4 (EM)
CD8 (EM)
Mock
Peptide mix
Mock
Peptide mix
T1
T2
IFNγ-PE
TNFα-BV785
0,02%
0,01%
0,17%
1,09%
0,52%
0,39%
0,10%
0,00%
0,08%
0,72%
0,10%
0,12%
0,03%
0,00%
0,07%
0,96%
0,36%
0,27%
0,07%
0,00%
0,13%
0,67%
0,01%
0,25%



**Supplementary Figure B-6. Features of immune response in patient S. A:** Gating strategy used to define CD4+ and CD8+ Tem cells; **B:** CD4 and CD8 T-cell responses to peptide mixture of SARS-Cov-2 proteins S, N, M, ORF3a and ORF7a. Representative flow cytometry plots showing the cytokine profiles of SARS-CoV-2-specific CD4 and CD8 effector memory T cells after the stimulation; **C:** Bar-plots representing the percentage of different cytokine-producing populations of SARS-CoV-2-specific CD4 and CD8 T cells after background subtraction (data from the mock-stimulated sample were subtracted from peptide-stimulated samples). Time points T1 and T2 correspond to August 20, 2020 and February 16, 2021 respectively. Stimulation with peptide mixture of SARS-Cov-2 proteins S, N, M, ORF3a and ORF7a provoked expansion of both SARS-CoV-2-specific CD4 and CD8 T cells; the CD4 T-cell response predominated over CD8, as usual for COVID-19 patients (Q. Zhang, Zmasek, and Godzik 2010; Habel et al. 2020). The figure is adapted from (Stanevich et al., 2023).

# Supplementary Tables B

## Supplementary Table B-1: The timeline of patient S survey and therapy.

| Event or procedure | Dates and results | Comment |
|---|---|---|
| Contact with patient A | April 10, 2020 - April 16, 2020 | Shared a ward in a hospital |
| SARS-CoV-2 PCR tests | In 2020: April 17 (+), April 30 (+), May 14 (-), May 19 (-), June 9 (-), July 14 (-), August 3 (+), August 5 (+), August 8 (+), August 11 (+), August 13 (+), August 17 (+, Ct 22), August 20 (+, Ct 19), August 21 (+), August 26 (+), August 27 (+), August 31 (+), September 2 (+, Ct 32), September 3 (+), September 8 (+), September 12 (-), September 16 (-), November 10 (-), December 16 (-);<br>In 2021: January 9 (+), January 11 (+), January 15 (+), January 19 (+, Ct = 21), January 22 (+, Ct = 24), January 22 (+, Ct = 31), February 1 (+, Ct = 24), February 8 (+, Ct = 30), February 16 (+, Ct = 19), February 19 (+, Ct = 29), March 1 (+, Ct = 32), March 10 (-), April 5 (-) | (+) - positive PCR;<br>(-) - negative PCR;<br>Ct - real time PCR cycle threshold when known.. |
| Periods of pneumonia | In 2020: June 6 - September 1;<br>In 2021: January 9 - February 1. | |
| ELISA for anti-S-SARS-CoV-2 IgG | In 2020:<br>August 17, 2020: negative (Cut-off-Index 0.54);<br>November 12, 2020: positive (Cut-off-Index 3.75);<br>December 15: ambiguous (Cut-off-Index 1.03). | Cut-off-index for ELISA: <0,8 – negative; 0,8 – 1,1 – ambiguous; >1 - positive. |
| VN assay | In 2020:<br>August 17: neutralizing antibodies not detected (titer <10);<br>November 12: neutralizing antibodies not detected (titer <10);<br>December 15: neutralizing antibodies not detected (titer <10). | |
| Sequenced samples | Patient A: April 19, 2020;<br>Patient S: In 2020: August 17, August 20;<br>In 2021: January 11, January 19, January 22, February 19. | The January 11, 2021 swab was obtained prior to convalescent plasma transfusion on the same date. |
| Blood samples | August 20, 2020; February 16, 2021 | |
| Chemotherapy | In 2020:<br>CHOP: January 5 - January 9; R-EPOCH: January 24 - January 29; R-EPOCH: February 13 - February 18; R-EPOCH: March 5 - March 10; R-ICE: April 5 - April 8; R-ICE: June 23 - June 27; R-ICE: July 24 - July 28; R-GemOx: October 30; GemOx: October 30; GemOx: November 14; ICE: December 5 - December 12; BEAM: December 21 - December 27. | Abbreviations of chemotherapy regimens are standardized for DBCL (“Cancer Therapy Adviser,” n.d.) |
| Autological transplantation of hematopoietic stem cells (Auto-HSCT) | December 28, 2020 | |
| Convalescent plasma | In 2021: January 11, January 15, January 18. | |

**Supplementary Table B-2**: **The list of mutations and their frequencies in sequencing reads obtained from patient S swab samples.** Only variants reaching 30% frequency at least in one of the samples are shown. The consensus variants (read frequency > 50%) are highlighted in blue, nonsynonymous nucleotide substitutions are in bold. NC (no coverage) indicates coverage depth less than 4 reads. “Selection” and “Trend Z” columns mark positions that experience positive selection and increase of corresponding changes in frequency, according to observablehq.com (Sergei Pond 2020) accessed on 31th March 2020.

| Gene | Nucleotide change | AA change | Aug 17 2020 | Aug 20 2020 | Jan 11 2021 | Jan 19 2021 | Jan 22 2021 | Feb 19 2021 | Selection | TrendZ |
|---|---|---|---|---|---|---|---|---|---|---|
| leader | C:676:T | leader:G137G | 0.000 | 0.000 | 0.992 | 0.990 | 0.998 | 1.000 | | |
| nsp2 | G:1312:A | nsp2:L169L | 0.000 | 0.000 | 0.998 | 0.998 | 0.995 | 0.999 | | |
| | C:1441:T | nsp2:G212G | 0.995 | 0.995 | 0.996 | 0.996 | 0.995 | 0.992 | | |
| | T:1552:C | nsp2:A249A | 0.000 | 0.000 | 0.999 | 0.996 | 0.998 | 0.996 | | |
| | **C:2037:T** | **nsp2:A411V** | **0.000** | **0.000** | **0.996** | **0.996** | **0.999** | **1.000** | + | |
| nsp3 | **A:4229:G** | **nsp3:T504A** | **0.434** | **0.730** | **0.000** | **0.000** | **0.000** | **0.000** | | |
| | **A:4229:C** | **nsp3:T504P** | **0.000** | **0.000** | **0.994** | **1.000** | **0.999** | **1.000** | | |
| | **G:5180:A** | **nsp3:D821N** | **1.000** | **1.000** | **1.000** | **1.000** | **1.000** | **NC** | | |
| | **C:7086:T** | **nsp3:T1456I** | **0.000** | **0.000** | **1.000** | **1.000** | **1.000** | **NC** | + | + |
| nsp4 | T:9091:C | nsp4:S179S | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.333 | | |
| | **C:9438:T** | **nsp4:T295I** | **0.995** | **1.000** | **0.997** | **1.000** | **0.999** | **1.000** | | |
| | **G:9497:A** | **nsp4:V315I** | **0.974** | **1.000** | **0.937** | **0.958** | **0.916** | **NC** | | + |
| 3C | G:10318:A | 3C:K88K | 0.014 | 0.460 | 0.000 | 0.000 | 0.000 | 0.000 | | |
| nsp6 | TG:11082:T | nsp6:del37 | 0.000 | 0.005 | 0.302 | 0.221 | 0.238 | 0.175 | | |
| | **G:11083:T** | **nsp6:L37F** | **0.006** | **0.007** | **0.621** | **0.743** | **0.685** | **0.810** | | |
| | **G:11804:A** | **nsp6:V278I** | **0.000** | **0.000** | **0.000** | **0.000** | **0.000** | **0.480** | | |
| nsp7 | **T:12015:G** | **nsp7:V58G** | **0.997** | **0.998** | **1.000** | **0.999** | **0.999** | **0.999** | | |
| nsp9 | A:12886:G | nsp9:T67T | 1.000 | 0.993 | 0.997 | 0.999 | 0.998 | 1.000 | | |
| RdRp | C:13620:T | RdRp:D60D | 0.453 | 0.222 | 0.000 | 0.000 | 0.000 | NC | | |
| | **A:13913:G** | **RdRp:N158S** | **0.000** | **0.000** | **1.000** | **0.998** | **0.999** | **NC** | | |
| | C:14625:T | RdRp:C395C | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.391 | | |
| | A:15456:G | RdRp:S672S | 0.000 | 0.000 | 1.000 | 0.997 | 1.000 | 1.000 | | |
| helicase | C:16575:T | helicase:D113D | 0.000 | 0.000 | 1.000 | 1.000 | 1.000 | NC | | |
| | A:17337:G | helicase:T367T | 0.000 | 0.000 | 1.000 | 0.999 | 0.996 | NC | | |
| endornase | **C:20234:T** | **endornase:P205L** | **0.000** | **0.000** | **1.000** | **1.000** | **1.000** | **1.000** | + | + |
| S | **C:21711:T** | **S:S50L** | **1.000** | **0.998** | **1.000** | **1.000** | **1.000** | **1.000** | | |
| | ATACATG:21764:A | S:del68_70 | 0.354 | 0.626 | 0.000 | 0.000 | 0.000 | 0.000 | + | + |
| | TTTTGGGTGTTTA:21981:T | S:del140_144 | 1.000 | 0.922 | 0.962 | 0.965 | 0.967 | 1.000 | + | + |
| | **G:22381:T** | **S:R273S** | **NC** | **0.000** | **1.000** | **NC** | **1.000** | **NC** | | |
| | C:22675:A | S:S371S | NC | 0.636 | 0.000 | NC | 0.000 | NC | | |
| | **T:22882:G** | **S:N440K** | **NC** | **0.000** | **NC** | **NC** | **1.000** | **NC** | + | + |
| | **T:22917:G** | **S:L452R** | **NC** | **0.000** | **NC** | **NC** | **1.000** | **NC** | | + |
| | **A:22920:T** | **S:Y453F** | **NC** | **0.625** | **NC** | **NC** | **0.000** | **NC** | | |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **C:22971:A** | **S:T470N** | **NC** | **0.992** | **NC** | **1.000** | **1.000** | **NC** | | |
| | **G:22988:A** | **S:G476S** | **NC** | **0.000** | **NC** | **1.000** | **1.000** | **NC** | + | |
| | **C:23423:T** | **S:P621S** | **0.349** | **0.322** | **0.000** | **0.000** | **0.000** | **0.000** | | |
| | **A:23772:G** | **S:D737G** | **0.999** | **1.000** | **0.991** | **0.999** | **1.000** | **0.989** | | |
| ORF3a | T:25435:C | ORF3a:L15L | 0.000 | 0.000 | 0.000 | 0.530 | 0.239 | 1.000 | | |
| | **C:26261:T** | **E:S6L** | **0.000** | **0.060** | **1.000** | **1.000** | **1.000** | **1.000** | | |
| | **T:26320:C** | **E:F26L** | **0.000** | **0.000** | **NC** | **1.000** | **1.000** | **NC** | | |
| E | T:26445:C | E:S67S | 1.000 | 1.000 | NC | 1.000 | 1.000 | NC | | |
| M | **T:26908:G** | **M:L129R** | **0.186** | **0.047** | **1.000** | **1.000** | **0.967** | **NC** | | |
| ORF6 | T:27351:C | ORF6:S50S | 0.000 | 0.000 | 0.968 | 1.000 | 1.000 | NC | | |
| ORF7a | GATT:27758:G | ORF7a:del2 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | NC | | |
| ORF8 | **C:27945:T** | **ORF8:Q18*** | **1.000** | **0.995** | **1.000** | **0.987** | **1.000** | **NC** | | |
| | **C:28289:A** | **N:P6T** | **0.140** | **0.061** | **0.993** | **0.993** | **0.997** | **0.960** | | |
| N | **A:28856:G** | **N:R195G** | **0.428** | **0.702** | **0.000** | **0.000** | **0.000** | **0.000** | + | |

**Supplementary Table B-3**. **Results of HLA calling from WES data using HLA-HD.** Classical HLA genes A, B, C, DRB1 and DQB1 were additionally confirmed with HLA genotyping (see Methods). These alleles were further combined into haplotypes via HLA-2-Haplo software. Thus, haplotype 1 corresponds to A - 01:01, B - 08:01, C - 07:01, DRB1 - 03:01, DQB1 - 02:01 (population frequency 5.98e-2); and haplotype 2 corresponds to A - 03:01, B - 07:02, C - 07:02, DRB1 - 01:01, DQB1 - 05:01 (population frequency 3.37e-3). Worldwide allele frequencies are presented according to (Solberg et al. 2008; Sarkizova et al. 2020).

| **HLA** | **Allele 1** | **Frequency (Allele 1)** | **Allele 2** | **Frequency (Allele 2)** |
|---|---|---|---|---|
| A | HLA-A*03:01:01 | 0.04272 | HLA-A*01:01:01 | 0.04843 |
| B | HLA-B*08:01:01 | 0.02960 | HLA-B*07:02:01 | 0.04104 |
| C | HLA-C*07:02:01 | 0.13101 | HLA-C*07:01:01 | 0.06887 |
| DRB1 | HLA-DRB1*03:01:01 | 0.0676 | HLA-DRB1*01:01:01 | 0.04123 |
| DQA1 | HLA-DQA1*01:01:01 | - | HLA-DQA1*05:01:01 | - |
| DQB1 | HLA-DQB1*05:01:01 | 0.09307 | HLA-DQB1*02:01:08 | 0.15003 |
| DPA1 | HLA-DPA1*01:03:01 | - | - | - |
| DPB1 | HLA-DPB1*04:02:01 | 0.18989 | HLA-DPB1*04:01:01 | 0.23267 |

**Supplementary Table 4**. **List of peptides experimentally validated in previous studies and included in the IEDB database that overlapped the mutations observed in this study.**

| Mutation | HLA class | Peptide before mutation | Peptide after mutation | HLA and change of percentile rank |
|---|---|---|---|---|
| S:S50L | HLA I | FRSSVLH**S**T | FRSSVLH**L**T | HLA-C*07:01: Weak (0.74) -> Weak (0.93) |
| S:S50L | HLA I | **S**TQDLFLPF | **L**TQDLFLPF | HLA-A*01:01: Weak (1.1) -> Weak (1.6) |
| S:S50L | HLA I | **S**TQDLFLPFF | **L**TQDLFLPFF | HLA-A*01:01: Weak (1.2) -> Weak (1.7) |
| S:del140_144 | HLA I | CNDP**FLGVY** | CNDPYHKNN | HLA-A*01:01: Strong (0.39) -> No binding |
| S:del140_144 | HLA I | **GVY**YHKNNK | KNNKSWMES | HLA-A*03:01: Strong (0.046) -> No binding |
| S:del140_144 | HLA I | FCNDP**FLGVY**Y | FCNDPYHKNNK | HLA-A*01:01: Weak (0.59) -> No binding |
| S:R273S | HLA I | YLQP**R**TFLL | YLQP**S**TFLL | HLA-B*08:01: Strong (0.019) -> Weak (0.69) |
| S:T470N | HLA I | KPFERDIS**T**EI | KPFERDIS**N**EI | HLA-B*07:02: Strong (0.11) -> Strong (0.15) |
| nsp3:T504A | HLA I | **T**DNYITTY | **A**DNYITTY | HLA-A*01:01: Strong (0.14) -> Weak (0.72) |
| nsp3:T504A | HLA I | P**T**DNYITTY | P**A**DNYITTY | HLA-A*01:01: Strong (0.007) -> Strong (0.07) |
| nsp3:T504P | HLA I | **T**DNYITTY | **P**DNYITTY | HLA-A*01:01: Strong (0.14) -> No binding |
| nsp3:T504P | HLA I | P**T**DNYITTY | P**P**DNYITTY | HLA-A*01:01: Strong (0.007) -> Weak (0.73) |
| nsp3:D821N | HLA I | TT**D**PSFLGRY | TT**N**PSFLGRY | HLA-A*01:01: Strong (0.001) -> Strong (0.068) |
| nsp3:D821N | HLA I | HTT**D**PSFLGRY | HTT**N**PSFLGRY | HLA-A*01:01: Strong (0.04) -> Strong (0.2) |
| nsp3:D821N | HLA I | TT**D**PSFLGRYM | TT**N**PSFLGRYM | HLA-A*01:01: Strong (0.089) -> Weak (1.2) |
| nsp3:T1456I | HLA I | S**T**NVTIATY | S**I**NVTIATY | HLA-A*01:01: Strong (0.097) -> Weak (0.67) |
| endornase:P205L | HLA I | K**P**RSQMEIDF | K**L**RSQMEIDF | HLA-B*07:02: Strong (0.3) -> No binding |
| ORF8:Q18* | HLA I | **QSCTQHQPY** | - | HLA-A*01:01: Strong (0.48) -> Lost |
| ORF8:Q18* | HLA I | **EPKLGSLVV** | - | HLA-B*07:02: Strong (0.49) -> Lost |
| ORF8:Q18* | HLA I | **VDDPCPIHFY** | - | HLA-A*01:01: Strong (0.16) -> Lost |
| N:P6T | HLA I | G**P**QNQRNAPRITF | G**T**QNQRNAPRITF | HLA-B*07:02: Strong (0.49) -> No binding |
| M:L129R | HLA I | VPLHGTI**L** | VPLHGTI**R** | HLA-B*07:02: Strong (0.29) -> No binding |
| ORF8:Q18* | HLA II | **PCPIHFYSKWYIRVG** | - | HLA-DRB1*01:01: Strong (0.54) -> Lost |

## Supplementary Notes B

### Supplementary Note B-1

Patient's BMI was 18.07, blood group - 0 (I), Rh+. Concomitant diagnoses did not include diabetes, cardiovascular diseases, anamnesis of stroke, acute myocardial infarction, thromboembolism, or chronic lung obstruction disease. Between June and September, 2020, there could be an invasive mycoses of lungs, but mycological assessment did not reveal any mycotic agents except *Candida* spp. There was no candidemia or bacteremia. Vancomycin-resistant Enterococcus (VCE) was detected in stool specimens during hospitalization in June-August 2020 and January 2021 but did not cause bacteremia. After bone marrow transplantation, no infection was detected by PCR, bacteriological or mycological expertise, or specimen's microscopy for infectious agents, besides VCE in stool. Patient S did not leave the Russian Federation between the beginning of 2020 and the end of SARS-CoV-2 infection, and had no pets at home during that period.

### Supplementary Note B-2

Between 19 and 47 genetic changes distinguish the patient S samples from the Wuhan-Hu-1/2019 reference strain (Wu et al. 2020). Seven of these changes, including the three SNPs at adjacent positions 21881-21883, were contained in each of the patient S samples, placing them in the B.1.1 lineage. The lineage of patient S carries the remaining 12 to 40 genetic changes. The patient A sample carries the seven mutations characteristic of B.1.1 but no other mutations, confirming patient A as the likely source of infection for patient S, and indicating that the remaining changes are specific to patient S.

### Supplementary Note B-3

Among the 12 mutations specific to patient S and observed in both August 2020 samples, 10 were single-nucleotide mutations (6 nonsynonymous, 3 synonymous and 1 creating a premature stop codon), and the remaining 2 were in-frame deletions. In the second August 2020 sample, 6 additional changes (4 nonsynonymous, 1 synonymous and 1 in-frame deletion) reached consensus frequencies.

Six of the mutations that reached consensus frequencies in the August 2020 samples reversed back to the ancestral state by January 2021, including the ΔF combination (see **Supplementary Note B-4**). Additionally, the January-February 2021 samples gained 21 new mutations compared to the August 17, 2020 sample. 10 of these mutations (6 nonsynonymous and 4 synonymous) were detected in all winter samples. The other 11 mutations (8 nonsynonymous and 3 synonymous) were each called in a subset of the winter samples; in the remaining samples, the corresponding sites were usually poorly covered. Overall, 34 changes were observed in the January 22, 2021 sample, which is the highest-quality sample among the winter 2021 samples (**Figure 4.1D**, **Supplementary Table B-2**). Together with the six reverted changes, this totals to 40 observed changes.

In addition to changes in the consensus sequence, we observed a number of variants at intermediate frequencies (above 30% in at least one of the samples, but below 50% in all samples and therefore not included in the consensus sequence; **Figure 4.1D**, **Supplementary Table B-2**), indicating within-host polymorphism. Three such variants (1 nonsynonymous and 2 synonymous) were observed in the August samples (all of them were lost in the January-February samples), and three (2 synonymous and 1 frame-disrupting deletion) were observed in the January-February samples (all absent in the August samples).

### Supplementary Note B-4

Among the positions that acquired amino acid mutations, ten (nsp2:A411V, nsp3:T1456I, nsp4:V315I, endornase:P205L, S:del68_70, S:del140_144,S:N440K, S:L452R, S:G476S, N:R195G) experienced pervasive positive selection according to the FEL (fixed effects likelihood) model (Kosakovsky Pond and Frost 2005) and/or their frequencies grew in the global viral population according to Jonckheere's trend test **(Supplementary Table B-2**), as reported in (Pond et al., 2020) (accessed on 31th March 2020).

Many of the detected mutations are known from other studies. Notably, these include the ΔF combination (S:Y453F + S:Δ69-70HV), which is observed in the consensus of the August 20, 2020 sample; S:Y453F is also found at high read frequency in the other 2020 sample, indicating that ΔF was probably also present at this time point (S:69-70 is too poorly covered at this time point to be called). The ΔF combination was previously described as associated with mink-related clusters. It has arisen in parallel in multiple mink populations; among humans, it was mainly found in cases traceable to minks (“Cluster 5”, or B.1.1.298), indicating reverse transmission (Oude Munnink et al. 2021). Despite the presence of the ΔF combination, patient S cannot be placed into the cluster 5 clade because cluster 5 is separated from B.1.1 by two additional mutations (those at positions 15656 and 25936) which are absent in patient S (**Supplementary Figure B-3**). Furthemore, the ΔF combination was not fixed in the August 2020 samples of patient S but segregated at an intermediate frequency (**Supplementary Table B-2**). Together, this indicates that the ΔF combination was acquired by patient S independently. It was not observed in any of the 2021 patient S samples, indicating that it had been reversed by that time.

The ΔF combination confers the ability to rapidly replicate to high titers and to evade recognition by neutralizing antibodies (Lassaunière et al. 2020), raising concerns that these mutations may affect vaccine efficiency. Y453F affects the receptor-binding domain (RBD), possibly increasing hACE2 binding (J. Chen et al. 2020; Starr et al. 2021). It allows immune escape from monoclonal antibodies and polyclonal sera; in particular, it has led to 57% escape from the REGN10933 monoclonal antibody, a component of FDA-approved Regeneron’s REGN-COV2 cocktail for treatment of COVID-19 patients, although it did not allow escape from the full cocktail of two antibodies (REGN10933+REGN10987) (Betrains et al. 2021). It was also shown to escape cellular immunity in HLA-A24-restricted patients (Truong et al. 2021).

The second mutation of the ΔF combination, S:Δ69-70HV, was recently shown to occur in a virus from another immunocompromised patient with COVID-19 (Kemp et al. 2020) (**Supplementary Figure B-7**). In that study, S:Δ69-70HV has been fixed during convalescent plasma therapy, suggesting antibody selection pressure, which is consistent with decreased virus sensitivity to neutralization with sera from recovered patients. However, patient S was not treated with convalescent plasma in 2020 and had no detectable neutralizing antibody response, suggesting that S:Δ69-70HV could have been favored by some other factor of selection. In patient S, both the S:Y453F and the S:Δ69-70HV mutations were polymorphic in 2020 and were lost by 2021, suggesting that this other factor may have been transient. The presence of the ΔF combination in the August 19, 2020 sample may underlie reduced sensitivity to neutralizing antibodies for this time point (**Supplementary Figure B-4**). Reacquired sensitivity to neutralizing antibodies by February 19, 2021 is also consistent with the loss of the ΔF combination by this time (**Supplementary Figure B-4**).

Besides S:Δ69-70HV, patient S has acquired six additional mutations that were also observed in other immunocompromised patients: S:S50L (McCarthy et al. 2021), S:N440K (Andrew Rambaut 2020), S:Δ141-144 (McCarthy et al. 2021; Khatamzas, Rehn, et al. 2021; Agerer et al. 2021; Dolton et al. 2021), nsp3:T504I (Khatamzas, Rehn, et al. 2021), nsp3:T295I (McCarthy et al. 2021) and nsp6:L37F (Andrew Rambaut 2020) (**Supplementary Figure B-7**). The most recurrent of these mutations, S:Δ141-144, was shown to lead to an escape from neutralizing antibodies (Wang et al. 2020). It falls into the recurrent deletion region (Wang et al. 2020) where frequent deletions are observed, including S:Y144del, the lineage defining deletion of B.1.1.7 (Focosi and Maggi 2021) which is speculated to have been founded by a chronically infected individual[1]). Another recurrent mutation, nsp6:L37F, is associated with asymptomatic course of infection (Pereira 2020); plausibly, it could have contributed to the ultimate recovery of the two immunocompromised patients in whom it has been observed (patient S and patient 3 from Truong et al., 2021).

Two of the mutations that emerged in patient S also spread in the general population as part of variants of concern (VOCs). The first is S:Δ69-70HV, which is a lineage-defining mutation of B.1.1.7. The second is S:L452R, which is found in several VOCs, including AY.1, AY.2 and B.1.617.2, as well as in multiple variants of interest. S:L452R was shown to have a pleiotropic effect, causing an escape both from T cell immunity and from neutralizing antibodies (Truong et al. 2021; Zinzula 2021). Finally, ORF8:Q18*, a stop-inducing mutation in ORF8, is reminiscent of the stop-inducing mutation in a different codon of the ORF8 protein (27th, as opposed to the 18th in this study) which is another of the lineage-defining mutations of B.1.1.7. The functions of ORF8 and its role in immune response and disease progression are extensively debated (Pereira 2020; Zinzula 2021; Y. Zhang et al. 2021, 8).

### Supplementary Note B-5

We excluded ORF8:Q18* from the analysis presented in **Figure 4.2** because it is impossible to calculate a presentation score (BR, PHBR or imBR) for a site lost from an amino acid sequence due to a stop gain. However, among the peptides absent from the ORF8 sequence after the

ORF8:Q18* change, three were listed in IEDB as immunogenic on HLA I alleles carried by patient S, and one, as immunogenic on an HLA II allele carried by patient S. All of them are strong binders (**Supplementary Table B-4**). Therefore, ORF8:Q18* also causes T cell escape.